STAFFING OF TIME-VARYING QUEUES TO ACHIEVE TIME-STABLE PERFORMANCE

UNABRIDGED VERSION: INTERNET SUPPLEMENT

by

Z. Feldman

Technion Institute
Haifa 32000
ISRAEL
zoharf@tx.technion.ac.il

A. Mandelbaum

Technion Institute
Haifa 32000
ISRAEL
avim@ie.technion.ac.il

W.A. Massey

Princeton University
Princeton, NJ 08544
U.S.A.
wmassey@princeton.edu

W. Whitt

Columbia University
New York, NY 10027-6699
U.S.A.
ww2040@columbia.edu

November 2004; Revision: November 27, 2005

Abstract

This is a longer version of a paper with the same title, which has been submitted to *Management Science*.

Abstract from the Journal Version

This paper develops methods to determine appropriate staffing levels in call centers and other many-server queueing systems with time-varying arrival rates. The goal is to achieve targeted time-stable performance, even in the presence of significant time-variation in the arrival rates. The main contribution is a flexible simulation-based *iterative-staffing algorithm* (ISA) for the $M_t/G/s_t + G$ model - with nonhomogeneous Poisson arrival process (the M_t) and customer abandonment (the $+G$). For Markovian $M_t/M/s_t + M$ special cases, the ISA is shown to converge. For that $M_t/M/s_t+M$ model, simulation experiments show that the ISA yields time-stable delay probabilities across a wide range of target delay probabilities. With ISA, other performance measures - such as agent utilizations, abandonment probabilities and average waiting times - are stable as well. The ISA staffing and performance agree closely with the modified-offered-load (MOL) approximation, which was previously shown to be an effective staffing algorithm without customer abandonment. While the ISA algorithm so far has only been extensively tested for $M_t/M/s_t + M$ models, it can be applied much more generally, to $M_t/G/s_t + G$ models and beyond.

What is Contained Here?

This longer version presents more examples; e.g., it treats the $M_t/M/s_t$ model (without customer abandonment) and the $M_t/M/s_t + M$ model with $\theta > \mu$ and $\theta < \mu$, where θ is the abandonment rate and μ is the service rate. It treats the challenging example from Jennings et al. (1996). There is extra detail for the previous examples; there are 47 figures here, but only 10 in the journal version. This longer version also provides additional theoretical support.

Contents

1. Introduction

Service systems such as banks, insurance companies and hospitals play an important role in our society. Services employ about 60–80% of the work force in western economies, and their importance is sharply on the rise, both within service and manufacturing companies. In our service-driven economy, it is estimated that over 70% of the business transactions are carried out over the phone. Most of these transactions are processed by telephone call centers, which have become the preferred and prevalent means for companies to communicate with their customers. For an overview of call centers and models of them, readers are referred to the recent review by Gans, Koole and Mandelbaum (2003).

The modern call center is a highly complex operation that fuses advanced technology and human beings. But the economic and managerial significance of the latter clearly outweighs the former. More specifically, labor costs (agents' salaries, training, etc.) typically run as high as 70% of the total operating costs of a call center, and attrition rates in call centers reach anywhere from 30% per year (considered low) to over 200% at times. In such circumstances, perhaps the most important operational decision to be made is staffing: what is the appropriate number of telephone agents that are to be accessible for serving calls. Overstaffing is wasteful, while understaffing leads to low service levels and overworked agents.

1.1. The Staffing Problem

The staffing problem typically takes the following form: Under an existing operational reality, and given a desired quality of service, we seek the least number of agents at each time that is required to meet a given service-level constraint. This problem, which has received much attention over the years (see Section 4 in Gans et. al.), is challenging both theoretically and practically. The challenges are easy to understand, because the natural model for the staffing problem is a many-server queue with a time-varying arrival rate, which is notoriously difficult to analyze. The practical importance of staffing is highlighted by considering a bank employing 10,000 telephone agents and catering to millions of customers per day; even small gains in operational efficiency or service quality clearly can provide great benefit.

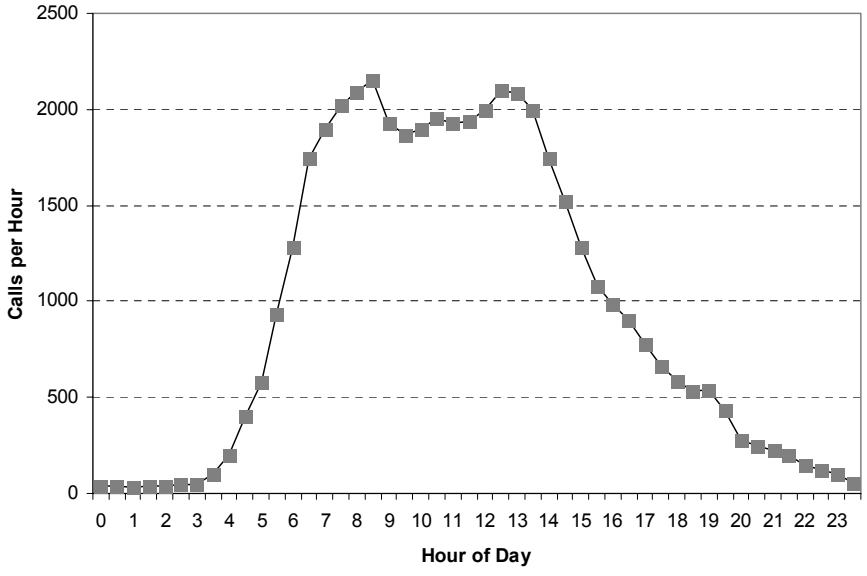
Figure 1 depicts a typical arrival-rate function to a telephone call center. Call volumes are low around midnight (hour 0), starting to increase in the early hours of the morning, peaking at late morning, then dropping somewhat around midday (12, lunch break), rising again afterwards, and then dropping thereafter to midnight levels. The displayed arrival-rate

function is an average of several similar days; the actual number of arrivals, in a given hour on a given day, fluctuates randomly around this average. (The functional form in Figure 1 is typical; the particular values for the arrival rates come from Green, Kolesar and Soares (2001).)

Staffing planners are thus faced with two sources of variability: **predictable variability** – time-variations of the expected load – and **stochastic variability** – random fluctuations around this time-dependent average. Most available staffing algorithms are designed to cope only with stochastic variability; they avoid the predictable variability in various ways. For example, when the service times are relatively short, as in many call centers when service is provided by a telephone call, it is usually reasonable to use a *pointwise stationary approximation* (PSA), i.e., to act as if the system at time t were in steady-state with the arrival rate occurring at that instant (or during that half hour). With PSA, one performs a stationary or steady-state analysis with a stationary model having parameters that vary by the time of day; see Green and Kolesar (1991) and Whitt (1991). (The PSA is the leading term in the more sophisticated *uniform-acceleration* (UA) approximation; see Massey and Whitt (1998) and references therein.)

However, service times are not always short, even in call centers. If relatively lengthy interactions are not uncommon, then PSA tends to be inappropriate. When service times are not so short, significant predictable variability can cause PSA to produce poor performance. As a consequence, some parts of the day may be overstaffed, while others are understaffed.

Figure 1: **Hourly call volumes to a medium-size call center**



For a review of staffing methods, see Green, Kolesar and Whitt (2005).

In this paper we address the staffing problem with *both* predictable and stochastic variability. Here is the problem we aim to solve: **Given a daily performance goal, and faced with both predictable and stochastic variability, we seek to find the minimal staffing levels that meet this performance goal stably over the day.** In particular, we aim to find an appropriate time-dependent staffing function for **any** arrival-rate function, where "appropriate" means that we achieve time-stable performance. We aim to agree with PSA when PSA is appropriate and do significantly better when it is not appropriate. We emphasize the importance of achieving time-stable performance. With time-stable performance, the nearly-constant quality of service is easily adjusted up or down, as desired.

The main contribution of this paper is a flexible simulation-based *iterative-staffing algorithm* (ISA). We develop the ISA for the many-server $M_t/G/s_t + G$ queueing model, which has a nonhomogeneous Poisson arrival process (the M_t) with time-varying arrival-rate function $\lambda(t)$, independent and identically distributed (i.i.d.) random service times with a general cumulative distribution function or cdf (the first G), a time-varying number of servers s_t, which is for us to set, and i.i.d. random times to abandon (before starting service) with a general cdf (the final $+G$). And we show that the ISA yields time-stable performance across a wide range of delay-probability targets for Markovian $M_t/M/s_t + M$ special cases (where the service-time and time-to-abandon cdf's G and F are exponential). The ISA uses a target delay probability, so stability is directly achieved for the delay probabilities, but other performance measures are quite stable as well.

Even though we only report results for ISA applied to Markovian $M_t/M/s_t + M$ models, the method is developed for more general $M_t/G/s_t + G$ models. (Indeed, we obtained similar results for log-normal and deterministic service-time distributions.) Moreover, the ISA applies much more generally, so that it has the potential of far-reaching applications. Indeed, by being based on simulation, ISA has two important advantages: First, by using simulation, we achieve *generality*: We can apply the approach to a large class of models; we are not limited to models that are analytically tractable. We are able to include realistic features, not ordinarily considered in analytical models. For example, we can carefully consider what happens to agents who are in the middle of a call when their scheduled shift ends. Second, by using simulation, we achieve *automatic validation*: In the process of performing the algorithm, we directly confirm that ISA achieves its goal; we directly observe the performance of the system under the final staffing function $\{s_t^{ISA} : 0 \le t \le T\}$. Of course, in other settings the effectiveness of the ISA

still needs to be verified.

Although we do not discuss many-server heavy-traffic stochastic-process limits here, they play a prominent role in supporting what we do. For example, they provide theoretical support for the fundamental square-root-staffing formula in (2.7) below. Moreover, this paper shows that many of the insights about the performance of stationary many-server Markovian queues provided by the many-server heavy-traffic stochastic-process limits in Halfin and Whitt (1981) and Garnett et al. (2002) carry over to corresponding queueing systems with time-varying arrivals, provided that proper staffing is done, aimed at achieving time-stable performance. With the ISA staffing, the global performance turns out to be essentially the same as for the stationary models, being well described by the heavy-traffic approximation formulas. That is dramatically demonstrated by Figures 11, 24, 32, 36 and 41 here.

1.2. Organization of this Paper

More than half this paper repeats the shorter journal version. We have added subsections and a Table of Contents to better communicate the organization.

The material in §§2-6 mostly repeats what is in the shorter main paper; there are only a few additions. Most of the additions are tables showing additional aspects of the performance of the ISA. We start in §2 by briefly reviewing the previous contributions by Jennings et al. (1996). We then overview our main contributions in this paper in §3. In §4 we specify our iterative-staffing algorithm in detail.

We start discussing simulation experiments in §5. In §5 we illustrate the performance of our algorithm by considering an Erlang-A-model ($M_t/M/s_t + M$) example (with abandonment). In §6 we analyze the realistic example, related to Figure 1, including short service times (in particular 6 minutes). In contrast to Green et. al. (2001), which is the source of Figure 1, we incorporate abandonment, which significantly impacts staffing results. When abandonment is present, as it often is in practice, it is possible to achieve good performance with significantly fewer agents than when abandonment is present. We show that conservative rules of thumb without abandonment tend to overstaff substantially.

We start presenting new results, not included in the main paper, here in §7. In §7, for comparison, we consider a time-avaying Erlang-C-model ($M_t/m/s_t$) example (without abandonment) and show that ISA again performs well. In §8 we revisit the "challenging example" in Jennings et al. (1996), which is again a $M_t/m/s_t$ model, now applying our iterative-staffing algorithm instead of the infinite-server and modified-offered-load (MOL) approximations used

before. We show that ISA performs essentially the same as MOL. In §9 we expand the analysis of the Erlang-A example from §5 by considering different patience parameters. In §5 we let the abandonment rate equal the service rate, which is often realistic in practice (approximately). However, in §9 we let the abandonment rate be 5 times and 0.2 times the service rate, representing the cases of very impatient customers and very patient customers, respectively. We show that the staffing methods continue to perform well in those alternative cases.

In §§10 and 11 we return to material in the journal version. In §10 we present some supporting theory. That mostly repeats what is in §6 of the journal version, but the final subsection does not appear there. In §11, we discuss the dynamics of the iterative algorithm, establishing monotonicity and convergence results. That too mostly repeats what is in the main paper (§7), but includes some additional figures and discussion.

The remaining material is not in the main paper: In §4.3, we explain and define the performance measures used in our simulations. In §12 we provide additional insight into the square-root-staffing formula from the perspective of many-server heavy-traffic limits, using the Markovian-service-network framework from Mandelbaum, Massey and Reiman (1998).

Finally, in §13 we summarize our main contributions and discuss directions for future research.

2. Our Point of Departure

Our point of departure is our (with Otis B. Jennings) previous paper: Jennings, Mandelbaum, Massey and Whitt (1996). There we considered the $M_t/G/s_t$ model (without customer abandonment), having a nonhomogeneous Poisson arrival process with arrival-rate function $\lambda(t)$ and *independent and identically distributed* (IID) service times $\{S_n : n \geq 1\}$, distributed as a random variable S with a general *cumulative distribution function* (cdf) G having mean $E[S] = 1/\mu$.

Let N_t be the number of customers in the $M_t/G/s_t$ system, either waiting or being served, at time t. We focused on the probability of delay, aiming to choose the time-dependent staffing level s_t such that

$$P(N_t \geq s_t) \leq \alpha < P(N_t \geq s_t - 1) \quad \text{for all} \quad t \ , \tag{2.1}$$

where α is the target delay probability.

In (2.1) above, we choose a constant target delay probability α for all times t. To achieve that for time varying arrival rate $\lambda(t)$, we aim to find an appropriate staffing function s_t.

This problem is challenging because the time-dependent delay probability $P(N_t \geq s_t)$ in (2.1) depends on the staffing function before time t as well as at time t.

In this section we review staffing algorithms based on *infinite-server* (IS) and *modified-offered-load* (MOL) approximations from Jennings et al. (1996). These approximations were developed for the $M_t/G/s_t$ model without customer abandonment, but the methods extend directly to the corresponding model with customer abandonment. The effectiveness of these methods with abandonments was not demonstrated previously, though. Our simulation experiments here will show that ISA produces essentially the same results as MOL, with and without customer abandonment, and that both are effective. (Our reported experiments are limited to Markovian $M_t/M/s_t + M$ models, but limited experimentation for other $M_t/G/s_t + G$ models indicate that excellent results hold there too.)

2.1. An Infinite-Server Approximation

We discuss the MOL and infinite-server approximations together, because the MOL approximation builds on the infinite-server approximation. We start by considering the infinite-server approximation. Why would anyone consider an infinite-server approximation? From a mathematical perspective, the reason is that the finite-server $M_t/G/s_t + G$ model of interest is analytically intractable, whereas the corresponding infinite-server $M_t/G/\infty$ model is remarkably tractable. From an engineering perspective, the reason is that the infinite-server model can be used to show the amount of capacity that would actually be used (and is thus needed) if there were no capacity constraints (i.e., a limited number of servers). For the Markovian $M_t/M/s_t + M$ model, where $\theta = \mu$, there is even a stronger connection: In that special case, the distribution of the number of customers in the infinite-server $M_t/M/\infty$ model actually *coincides* with the distribution of the number of customers in the $M_t/M/s_t + M$ model, as we explain in §10, so there is additional strong motivation for considering the infinite-server approximation.

So what does the infinite-server approximation do? The infinite-server approximation for the $M_t/G/s_t + G$ model approximates the random variable N_t by the number N_t^∞ of busy servers in the associated $M_t/G/\infty$ model, having infinitely many servers but the same arrival process and service times. *The infinite-server staffing function s_t^∞ is obtained by applying* (2.1) *with N_t^∞ instead of N_t.* That approximation provides great simplification because (i) the tail probability $P(N_t^\infty \geq s_t)$ at time t depends on the staffing function $\{s_t : t \geq 0\}$ only through its value at the single time t and (ii) the exact time-dependent distribution of N_t^∞ is known.

The first simplification follows from the fact that the distribution of the stochastic process $\{N_t^\infty : t \geq 0\}$ is totally independent of the staffing function $\{s_t : t \geq 0\}$. When we calculate $P(N_t^\infty \geq s_t)$, the staffing level s_t just serves as the argument of the tail-probability function. The second simplification stems from basic properties of $M_t/G/\infty$ queues. In particular, as reviewed in Eick et al. (1993a), for each t, N_t^∞ has a *Poisson distribution* whenever the number in the system at the initial time has a Poisson distribution. (Being empty is a degenerate case of a Poisson distribution.) That Poisson distribution is fully characterized by its mean m_t^∞.

The Offered Load. As in previous work, such as Eick et al. (1993a,b) and Jennings et al. (1996), our work reported here shows that the time-dependent mean m_t^∞ is the crucial quantity. We regard this exact time-dependent mean m_t^∞ in the $M_t/G/\infty$ model as the (*time-dependent*) *offered load* for the $M_t/G/s_t + G$ model.

We now observe that convenient formulas exist for the offered load m_t^∞. Eick et al. (1993a) showed that the offered load has the tractable representation

$$m_t^\infty \equiv E\left[N_t^\infty\right] = \int_{-\infty}^{t} G^c(t-u)\lambda(u)\,du = E\left[\int_{t-S}^{t} \lambda(u)\,du\right] = E\left[\lambda(t-S_e)\right]E[S]\ , \qquad (2.2)$$

where $\lambda(t)$ is the arrival-rate function, S is a generic service time with cdf G, $G^c(t) \equiv 1-G(t) \equiv P(S > t)$, and S_e is a random variable with the associated *stationary-excess cdf* (or equilibrium-residual-lifetime cdf) G_e associated with the service-time cdf G, defined by

$$G_e(t) \equiv P(S_e \leq t) \equiv \frac{1}{E[S]}\int_0^t G^c(u)\,du, \quad t \geq 0\ , \qquad (2.3)$$

with k^{th} moment $E[S_e^k] = E[S^{k+1}]/((k+1)E[S])$; see Theorem 1 of Eick et al. (1993a) and references therein.

The different expressions in (2.2) provide useful insight; see Eick et al. (1993a, b) and Section 4.2 of Green et al. (2005). For the special case in which $\lambda(t)$ is constant, $m_t^\infty \equiv m^\infty = \lambda E[S]$. Accordingly, the PSA approximation for m_t^∞ in the $M_t/G/\infty$ model is $m_t^{PSA} \equiv \lambda(t)E[S]$. We call m_t^{PSA} the PSA (time-dependent) offered load for the $M_t/G/s_t + G$ model.

In addition, there are convenient explicit formulas for m_t^∞ in special cases as well as useful approximations. We will use the explicit formula for sinusoidal arrival-rate functions in §5. Based on a second-order Taylor-series approximation for λ about t, the offered load can be approximated by

$$m_t^\infty \approx \lambda(t - E[S_e])E[S] + \frac{\lambda^{(2)}(t)}{2}Var(S_e)E[S]\ , \qquad (2.4)$$

where $\lambda^{(2)}(t)$ is the second derivative of the function λ evaluated at time t; see Theorem 9 of Eick et al. (1993a). Approximation (2.4) shows that the approximate offered load in (2.4)

coincides with the PSA offered load $m_t^{PSA} \equiv \lambda(t)E[S]$ except for a *time shift* by $E[S_e]$ and a *space shift* by $\lambda^{(2)}(t)Var(S_e)E[S]/2$. Since $\lambda^{(2)}(t)$ will be negative at a peak, we see that the actual requirements at times of peak demand are less than predicted by PSA. The mapping of the arrival-rate function into the infinite-server mean m_t acts as a smoothing operator, making the results less extreme. Of course, that is convenient for meeting practical constraints on staffing schedules (tours of duty).

The time shift is especially important. A simple refinemnt of PSA based on (2.4) suggested by Eick et al. (1993a) is *lagged PSA*, where we ignore the space shift and approximate m_t^∞ by $\lambda(t - E[S_e])E[S]$.

Since the infinite-server approximation suggests shifting from the PSA offered load $m_t^{PSA} \equiv \lambda(t)E[S]$ to the "infinite-server" offered load m_t^∞, it is useful to quantify the difference between these quantities. From (a special case of) Theorem 10 in Eick et al. (1993a), we can quantify the difference between the offered load m_t^∞ and the PSA offered load $m_t^{PSA} \equiv \lambda(t) \cdot E[S]$ in another way. Letting $(S_e)_e$ be a random variable with the twofold stationary-excess cdf $(G_e)_e$ (obtained by applying the stationary-excess operator twice to the service-time cdf G), we have the formula

$$m_t^\infty - \lambda(t) \cdot E[S] = E\left[\lambda'\left(t - (S_e)_e\right)\right] \cdot E[S_e] \cdot E[S] = \frac{1}{2} \cdot E\left[\lambda'\left(t - (S_e)_e\right)\right] \cdot E[S^2]. \qquad (2.5)$$

From (2.5), it follows that the PSA offered load will *not* be a good approximation of the infinite-server offered load when the arrival rate varies rapidly in time (large derivative λ'). For a given mean service time, they may also be far apart when the second moment of the service time, $E[S^2]$, (or variance) is large. The second condition has implications for non-exponential distributions that are heavy tailed; see Whitt (2000) for background.

The Normal Approximation. We now continue, exploiting the established Poisson distribution with a known time-dependent mean m_t^∞. Assuming that m_t^∞ is not extremely small, we can apply a *normal approximation* for the Poisson distribution, obtaining first $P(N_t \ge s_t) \approx P(N_t^\infty \ge s_t)$ and then

$$P(N_t^\infty \ge s_t) \approx P(N(m_t^\infty, m_t^\infty) \ge s_t) = P\left(N(0,1) \ge \frac{s_t - m_t^\infty}{\sqrt{m_t^\infty}}\right) = 1 - \Phi\left(\frac{s_t - m_t^\infty}{\sqrt{m_t^\infty}}\right), \qquad (2.6)$$

where $N(m, \sigma^2)$ denotes a normally distributed random variable with mean m and variance σ^2, and $\Phi(x) \equiv P(N(0,1) \le x)$ is the standard normal cdf.

The Square-Root-Staffing Formula. From (2.6), we see that we can obtain a stable approximate delay probability if we can choose the staffing function s_t^∞ to make $(s_t^\infty - m_t^\infty)/\sqrt{m_t^\infty}$ stable in the final term of (2.6). Accordingly, we obtain the *square-root-staffing formula*:

$$s_t^\infty = \lceil m_t^\infty + \beta\sqrt{m_t^\infty} \rceil, \quad 0 \le t \le T, \tag{2.7}$$

where $\lceil x \rceil$ is the least integer greater than or equal to x and the constant β is a measure of the *quality of service*. Combining the target in (2.1) and the normal approximation in (2.6), we see that the quality-of-service parameter β in (2.7) should be chosen so that $1 - \Phi(\beta) = \alpha$.

The normal approximation and the square-root-staffing formula for *stationary* many-server queues are classic results, see Whitt (1992) and references therein. What is less well understood is the role of the offered load m_t^∞ with time-varying arrivals. The notation s_t^∞ means that we staff according to the infinite-server approximation. In doing so, we not only apply the normal approximation and the square-root-staffing formula, but we also use the infinite-server mean m_t^∞ as the offered load.

2.2. The Modified-Offered-Load Approximation as a Refinement

Section 4 of Jennings et al. (1996) also introduced a refinement of the infinite-server approximation for the time-dependent delay probabilities, which is tantamount to a *modified-offered-load* (MOL) approximation, as in Jagerman (1975) and Massey and Whitt (1994, 1997). The MOL approximation for N_t in the $M_t/G/s_t + G$ model at time t, denoted by N_t^{MOL}, is the limiting steady-state number of customers in the system in the corresponding stationary $M/G/s + G$ model (with the same service-time and time-to-abandon distributions and the same number of servers s_t at time t), but using m_t^∞ as the stationary offered load operating at time t. Since the stationary offered load is $\lambda E[S]$, that means letting the homogeneous Poisson arrival process in the stationary $M/G/s + G$ model have time-dependent arrival rate

$$\lambda_t^{MOL} \equiv \frac{m_t^\infty}{E[S]} = m_t^\infty \mu \quad \text{at time} \quad t \ . \tag{2.8}$$

The MOL staffing function s_t^{MOL} is obtained by applying (2.1) *with N_t^{MOL} instead of N_t.*

The important insight is that the "right" time-dependent offered load in the $M_t/G/s_t + G$ model should be the time-dependent mean number of busy servers in the associated infinite-server model - m_t^∞. Since the right offered load for the stationary model is $\lambda E[S]$, the "obvious" direct time-dependent generalization is the PSA offered load $m_t^{PSA} \equiv \lambda(t)E[S]$. However, $\lambda E[S]$ is also the mean number of busy servers in the associated stationary infinite-server

model. It turns out that the mean number of busy servers in the infinite-server model is a better generalization of "offered load" than the PSA time-dependent offered load for most time-varying many-server models. Indeed, it may be considered exactly the right definition for the infinite-server model itself.

The MOL approximation in §4 of Jennings et al. (1996) was not applied directly. Instead of calculating the steady-state delay probability for the stationary $M/M/s$ model, we exploited an approximation for the delay probability based on a many-server heavy-traffic limit in Halfin and Whitt (1981). That produces a simple formula relating the delay probability α and the service quality β. Moreover, the heavy-traffic limit provides an alternative derivation of the square-root staffing formula in (2.7), without relying on an infinite-server approximation or a normal approximation. We will do the same thing here with customer abandonments, relying on the heavy-traffic limits for the $M/M/s+M$ model established by Garnett et al. (2002).

2.3. It Is Possible to Achieve Time-Stable Performance!

Jennings et al. (1996) showed that the method for setting staffing requirements in the $M_t/G/s_t$ model outlined above is remarkably effective. This was demonstrated by doing numerical comparisons for the $M_t/M/s_t$ special case. For any given staffing function, the time-dependent distribution of N_t in that Markovian model can be derived by solving a system of time-dependent ordinary differential equations (ODE's). We too could have exploited ODE's for the $M_t/M/s_t+M$ model, but we wanted to develop a method that applies to much more general models.

The most important conclusion from those previous experiments in Jennings et al. (1996) is that it is indeed possible to achieve time-stable performance for the $M_t/M/s_t$ model by an appropriate choice of a staffing function s_t, even in the face of a strongly time-varying arrival-rate function. Here we show the same is true for the $M_t/M/s_t+M$ model. And we provide a means to go far beyond these Markovian models.

3. Our Contributions Here

We develop staffing algorithms for more complicated time-varying many-server models, such as many-server queues with abandonment. For example, we treat the much more realistic $M_t/G/s+G$ model with non-exponential service times (the first G) and non-exponential abandonments (the $+G$). Allowing non-exponential service-time and time-to-abandon distributions is important, because they have been found to occur in practice; see Bolotin (1994) and Brown

et al. (2005). We emphasize that models with customer abandonment were not considered by Jennings et al. (1996) or anybody else (as far as we know).

For call centers, our ultimate goal is to treat realistic multi-server systems with multiple call types and skill-based routing (SBR), but we do not pursue that here. In that setting, it is natural to apply SBR methods for stationary models after using the MOL approximation in (2.8) for each call type at time t. Once we have reduced the problem to a stationary SBR model, we may be able to apply the staffing method in Wallace and Whitt (2005). Approaches based on these ideas remain to be investigated, however.

3.1. A Simulation-Based Iterative Staffing Algorithm (ISA)

Our first contribution is a simulation-based iterative-staffing algorithm (ISA) for many-server queues with time-varying arrival rate. By being based on simulation, ISA has two important advantages: First, by using simulation, we achieve **generality**: We can apply the approach to a large class of models; we are not restricted by having to have a model that is analytically tractable. We are able to include realistic features, not ordinarily considered in analytical models. For example, we can carefully consider what happens to agents who are in the middle of a call when their scheduled shift ends. Second, by using simulation, we achieve **automatic validation**: In the process of performing the algorithm, we directly confirm that ISA achieves its goal; we directly observe the performance of the system under the final staffing function $\{s_t : 0 \le t \le T\}$.

Following Jennings et. al. (1996), we assume that, in principle, any number of servers can be assigned at any time. In our implementation, however, time is divided into short intervals (we take 0.1 service times), and we keep the number of servers fixed over each of these small intervals. The service discipline is FCFS, and servers follow an exhaustive service discipline: a server that finishes a shift in the middle of a service will complete the service and sign out only when finished. (Our results prevail also for preemptive service disciplines under which servers leave at end-of-shifts and their customers, if any, are moved to the front of the queue; e.g., see Ingolfsson (2005).)

In practice, staffing is required to be fixed over longer staffing intervals - typically ranging from 15 minutes to an hour. Here we ignore that constraint. An initial staffing function with such constraints is obtained from our results by using in each staffing interval the maximum required staffing level at any time point within that staffing interval. That will yield an upper bound on the required staffing. Simulation can then be used, in the manner of the ISA, to

see if these initial staffing levels can be decreased, while still meeting the performance target at every time. See Green et al. (2005) for additional discussion on this point, and references studies of the impact of staffing intervals.

Continuing to follow Jennings et al. (1996), we use **the delay probability as our target performance measure**, but the same method could be applied to other performance measures. Specifically, given a target probability of delay, we identify time-varying staffing levels under which the actual probability of delay remains approximately equal to the given target at all times. Other performance measures, such as the average waiting time and the queue-length tail delay-probabilities, turn out to be relatively constant over time as well.

For the main model we study, the Markovian $M_t/M/s_t + M$ model, we not only implement and evaluate ISA, but we also provide a proof of convergence. To do so, we must set aside the (important) issue of estimating the time-dependent delay probability for any given staffing function by computer simulation, which is subject to statistical sampling error. That statistical sampling error decreases as we increase the number of independent replications, so it can be made arbitrarily small at the expense of computational effort, but for any given amount of computational effort it is always present. However, if we assume that we actually know the true delay probabilities associated with each staffing function, then we establish convergence. That is accomplished by applying sample-path stochastic-order notions, as in Whitt (1981).

3.2. An Extension of the Square-Root-Staffing Formula

While working with ISA, we discovered that the simulation-based solutions have astonishing regularity. In particular, we found that global performance measures coincide with the performance measures of the associated stationary model. In particular, when we used ISA to staff the time-varying $M_t/M/s_t + M$ model, we found that the staffing could be related to the steady-state behavior of the associated stationary $M/M/s + M$ model. That implies that the modified-offered-load approximation will work well for the $M_t/M/s_t + M$ model.

That leads us to our second contribution: We extend the **square-root staffing formula** based on the **modified-offered-load approximation** in Jennings et al. to the $M_t/M/s_t + M$ model. In particular, we suggest staffing according to the square-root-staffing formula in (2.7), where the service quality $\beta \equiv \beta(\alpha)$ is derived from a theoretical **one-to-one relation between α and β for the corresponding stationary model**.

In particular, we propose using $\beta(\alpha)$, for which staffing levels of $s = m + \beta\sqrt{m}$ would lead to the desired delay probability α in the corresponding stationary model. For that purpose, we

could calculate the steady-state probability of delay in the associated stationary $m/M/s+M$ model, which is routine because the number in system L_t is a birth-and-death stochastic process. We can also calculate all other performance measures in the $M/M/s+M$ model; e.g., see Garnett et al. (2002) and Whitt (2005).

However, instead of calculating the exact steady-state delay probability in the stationary model, we propose using an approximation for the steady-state delay probability - a simple formula based on a heavy-traffic limit - just as Jennings et al. applied the many-server heavy-traffic limit from Halfin and Whitt (1981). For the $M_t/M/s_t+M$ model, we use explicit formulas relating α to β obtained from the many-server heavy-traffic limits in Garnett, Mandelbaum and Reiman (2002).

We justify this simple analytic staffing formula by conducting experiments for the $M_t/M/s_t+M$ model, but we propose the approximation more generally. The effectiveness in any other context can be verified by applying the simulation-based ISA.

3.3. Staffing at the Offered Load

Finally, we make yet one more contribution. To describe it, we remind readers of the three heavy-traffic regimes for many-server queues: *Quality-Driven* (QD, lightly loaded), *Efficiency-Driven* (ED, heavily loaded) and *Quality-and-Efficiency-Driven* (QED, normally loaded); see Garnett et al. (2002). In our experiments for the many-server queue with abandonments we found that **simply staffing according to the offered load itself** is remarkably effective in the QED regime (specifically, where $\alpha = 0.5$), i.e., staffing by letting $s_t = m_t^\infty$ for the $M_t/M/s_t+M$ model works very well in the QED regime. Needless to say, abandonments play a crucial role in this property. This is another example of the importance of including abandonments in the model, when customers actually do abandon; see Garnett et al. (2002) for more discussion. See Section 12 for additional theoretical support, based on the Markovian-service-network framework of Mandelbaum, Massey and Reiman (1998).

3.4. The Naive Deterministic Approximation

Even though staffing according to the offered load is a remarkably simple method, there remains substantial sophistication, because we have to know that we should use the deterministic offered-load function m_t^∞. When the service times are relatively short (compared to the fluctuations in the arrival-rate function), we can use a truly **naive deterministic approximation**: We can then simply staff according to the PSA offered load: we can set $s_t^{ND} = m_t^{PSA} \equiv \lambda(t)/\mu$

(which will coincide with the offered load, m_t^∞, in that scenario). When we staff according to the PSA offered load $m_t^{PSA} \equiv \lambda(t)/\mu$, we are truly ignoring all stochastic variability; we are using only deterministic data about the model: the deterministic arrival-rate function $\lambda(t)$ and the deterministic mean service time $1/\mu$. Even though the infinite-server offered load m_t^∞ is a deterministic function, it depends on the service-time distribution beyond its mean, as is apparent from (2.2).

We conclude by mentioning that the naive deterministic approximation and PSA are actually not so effective in the setting of the realistic large example in Figure 1, when there is customer abandonment in the QED regime; see §6. Even though the service times are short for this realistic example, in particular, the mean service time is 6 minutes, the arrival-rate function changes very rapidly, especially in the hours $4 - 6$.

4. The Simulation-Based Iterative-Staffing Algorithm (ISA)

In this section we describe the simulation-based interactive-staffing algorithm (ISA). As indicated before, we determine time-dependent staffing levels aiming to achieve a given constant probability of delay at all times. In the process of applying the ISA, we directly confirm that our goal is being met. Indeed, the goal will necessarily be met, to a specified tolerance, if the algorithm converges. We then can confirm that other performance measures, such as server utilization, tail probabilities, average waits and abandonment probabilities, remain relatively stable as well.

4.1. The Simulation Framework

For our implementation of the algorithm, we assume that we have an $M_t/G/s_t+G \equiv M_t/GI/s_t+GI$ model with independent sequences of IID service times and IID times to abandon, which are independent of the arrival process, having general distributions, and a nonhomogeneous Poisson arrival process, which is fully specified by its arrival-rate function $\{\lambda(t); 0 \le t \le T\}$. (It will be evident that our approach extends to more general models.) For application of our algorithm, assuming that we use the $M_t/G/s_t + G$ model, there are issues about model fitting. For discussion about fitting non-homogeneous Poisson arrival processes, see Massey, Parker and Whitt (1996).

To start, we fix an arrival-rate function, a service-time distribution, a time-to-abandon (patience) distribution (when relevant) and a time-horizon $[0, T]$. For any random quantity of interest, let X_t^n denote the value at time t in the n^{th} iteration, for $t \in [0, T]$ (the given

time horizon). Although our algorithm is time-continuous, we make staffing changes only at discrete times. That is achieved by dividing the time-horizon into small intervals of length Δ. In all experiments presented in this paper, we use $\Delta = 0.1/\mu$, where $1/\mu$ is the mean service time. We then let the number of servers be constant within each of these intervals.

For any specified staffing function, the system simulation can be performed in a conventional manner. We generate a continuous-time sample path for the number in system by successively advancing the next generated event. The candidate next events are of course arrivals, service completions, abandonments and ends of shifts (the times at which the staffing function is allowed to change). For non-stationary Poisson arrival process, we can generate arrival times by thinning a single Poisson process with a constant rate λ^* exceeding the maximum of the arrival-rate function $\lambda(t)$ for all t, $0 \le t \le T$. Then an event in the Poisson process at time t (a potential arrival time) is in an actual arrival in the system with probability $\lambda(t)/\lambda^*$, independent of the history up to that time; see Section 5.5 of Ross (1990). Alternatively, the times between successive arrivals can be generated as independent events, according to probability distributions, determined by the last customer arrival time, and adjusted if necessary at ends of shifts.

In this section, let $s_t^{(n)}$ be the staffing level at time t in iteration n for $0 \le t \le T$. Let $N_t^{(n)}$ denote the random total number of customers in the system at time t, under this staffing function. We estimate the distribution of $N_t^{(n)}$ for each n and t by performing multiple (5000) independent replications. We think of starting off with infinitely many servers. Since this is a simulation, we choose a large finite number, ensuring that the probability of delay (i.e., of having all servers busy upon arrival) is negligible for all t. For the examples in §5 and §7, it suffices to let $s_t^{(0)} = 200$ for all t.

4.2. The Algorithm

The algorithm iteratively performs the following steps, until convergence is obtained. Here, convergence means that the staffing levels do not change much after an iteration. (Practically, they are allowed to change by some threshold τ, which we take to be 1.)

1. Given the i^{th} staffing function $\{s_t^{(i)} : 0 \le t \le T\}$, evaluate the distribution of $N_t^{(i)}$, for all t, using simulation.

2. For each t, $0 \le t \le T$, let $s_t^{(i+1)}$ be the least number of servers such that the delay-

probability constraint is met at time t; i.e., let

$$s_t^{(i+1)} = \arg\min \{k \in \mathbb{N} : P\left(N_t^{(i)} \geq k\right) \leq \alpha\} \ .$$

3. If there is negligible change in the staffing from iteration i to iteration $i+1$, then stop; i.e., if

$$\|s^{(i+1)} - s^{(i)}\|_\infty \equiv \max\{|s_t^{(i+1)} - s_t^{(i)}| : 0 \leq t \leq T\} \leq \tau \ ,$$

then stop and let $s^{(i+1)}$ be the proposed staffing function. Otherwise, advance to the next iteration, i.e., replace i by $i+1$ and go back to step 1. (We let $\tau = 1$.) ∎

For further discussion, let a superscript ISA denote the final iteration of ISA, so that s_t^{ISA} denotes the final staffing level at time t and N_t^{ISA} denotes the (random) number in system at time t with that staffing function s^{ISA}. Then, if the algorithm converges, it converges to a staffing function s^{ISA} for which $P\left(N_t^{ISA} \geq s_t^{ISA}\right) \approx \alpha$, $0 \leq t \leq T$.

Our implementation of ISA was written in C++. For the special case of the Markovian $M_t/M/s_t + M$ model, we rigorously establish convergence of the algorithm, as we explain in §11. Experience indicates that the algorithm consistently converges relatively rapidly. The number of iterations required depends on the parameters, especially the ratio $r \equiv \theta/\mu$, where θ is the individual abandonment rate. If $r = 1$, corresponding to an infinite-server queue (§10), then no more than two iterations are needed, since the distribution of the number in system does not depend upon the number of servers. As r departs from 1, the number of required iterations typically increases. For example, when $r = 10$, the number of iterations can get as high as $6 - 12$. When r is very small and the traffic intensity is very high, so that we are at the edge of stability, the number of iterations can be very large. For more discussion, see §11.

4.3. Estimating the Performance Measures

Throughout this paper we present several performance measures. Since we have time-varying arrivals, care is needed in their definition and estimation. In this subsection we describe our estimation procedure.

Most measures are time-varying. We define them for each time-interval t, and graph their values as function over $t \in [0, T]$. Other measures are global. They are calculated either as total counts (e.g. fraction abandoning during $[0, T]$), or via time-averages. We used $T = 24$ in all our simulations.

For replication k, the **delay probability** *in interval* t is estimated by the fraction of customers who are not served immediately upon arrival, out of all arriving customers during

the t time-interval. Namely, for the k^{th} replication, the estimator is:

$$\hat{\alpha}_k(t) = \frac{\sum_i 1\{customer\_i\_entered\_queue\_at\_interval\_t\}}{\sum_i 1\{customer\_i\_entered\_system\_at\_interval\_t\}} \equiv \frac{\hat{Q}_k(t)}{\hat{S}_k(t)} \ . \tag{4.1}$$

We obtain the overall estimator $\hat{\alpha}(t)$ by averaging over all replications. That was found to be essentially the same as (identical to for our purposes) the ratio of the average of $\hat{Q}_k(t)$ over all replications to the average of $\hat{S}_k(t)$.

For replication k, the estimator of the **average waiting time** *in interval* t is defined in an analogous way by

$$\hat{w}_k(t) = \frac{\sum_i w_i 1\{customer\_i\_entered\_system\_at\_interval\_t\}}{\sum_i 1\{customer\_i\_entered\_system\_at\_interval\_t\}} \tag{4.2}$$

where w_i is the total waiting time of customer i. Again we obtain the overall estimator $\hat{w}(t)$ by averaging over all replications.

The **average queue length** *in interval* t is taken to be constant over the time-interval. The queue length is also averaged over all replications. By the **tail probability** *in interval* t we mean specifically the probability that queue size is greater than or equal to 5 (taking 5 to be illustrative). Specifically, the indicators $1\{L_t^{(\infty)} - s_t^{(\infty)} \geq 5\}$ are averaged over all replications. (Here $L_t^{(\infty)}$ and $s_t^{(\infty)}$ are the values at time t obtained from the last iteration of ISA.)

For replication k, the estimator of the **server utilization** *in interval* t is the fraction of busy-servers during the time-interval, accounting for servers who are busy only a fraction of the interval:

$$\hat{\rho}_k(t) = \frac{\sum_{i=1}^{s_t^{(\infty)}} b_i}{s_t^{(\infty)} \cdot \Delta} \tag{4.3}$$

where b_i denotes the busy time of server i in interval t. Again, we obtain the overall estimator $\hat{\rho}(t)$ by averaging over all replications.

5. An Example with the Time-Varying Erlang-A Model

We demonstrate the performance of ISA by considering a time-varying Erlang-A model ($M_t/M/s_t+M$) with a special structured arrival-rate function.

5.1. A Sinusoidal Arrival-Rate Function

We consider a sinusoidal arrival-rate function. In particular, let the queueing system be faced with a non-homogeneous Poisson arrival process with a **sinusoidal arrival-rate function**

$$\lambda(t) = a + b \cdot \sin(ct), \quad 0 \leq t \leq T \ , \tag{5.1}$$

where $a = 100$, $b = 20$ and $c = 1$. Let the service times and the customer times to abandon (if they have not yet started service) come from independent sequences of independent and identically distributed (IID) exponential random variables, both having mean 1. As can be seen from PSA, the arrival rate is sufficiently large, that about 100 servers are required, so this example captures the many-server spirit of a call center. However, the sinusoidal form of the arrival-rate function is clearly a mathematical abstraction, which has the essential property of producing significant fluctuations over time, i.e., significant predictable variability. This particular arrival-rate function is by no means critical for our analysis; our methods apply to arbitrary arrival-rate functions, such as in Figure 1. (Indeed, for that, see Section 6.

An important issue, however, is the rate of fluctuation in the arrival-rate function compared to the expected service-time distribution. To be concrete, we will measure time in hours, and focus on a 24-hour day, so that $T = 24$. A cycle of the sinusoidal arrival-rate function in (5.1) is $2\pi/c$; since we have set $c = 1$, a cycle is $2\pi \approx 6.3$ hours. Thus there will be about 4 cycles during the day. That roughly matches the daily cycle in Figure 1 for the six-hour period around 12:00 noon.

Since we let the mean service time be 1 and have chosen to measure time in hours, the mean service time in this example is 1 hour. That clearly is relatively long for most call centers, where the interactions are short telephone calls. If we were to change the time units in order to rectify that, making the expected service time 10 minutes, then a cycle of the arrival-rate function would become about 1 hour, making for more rapid fluctuations in the arrival rate than are normally encountered in call centers. Thus our example is more challenging than usually encountered in call centers, but may be approached in evolving contact centers if many interactions do indeed take an hour or more. (We consider a practical example directly related to Figure 1 in §6.) From this preliminary analysis, we should anticipate that the service times are sufficiently long in our example that the traditional PSA method is likely to perform poorly here, just as in Jennings et al. (1996), and it does. As before, we are deliberately choosing a difficult case.

The arrival rate coincides with the PSA offered load, because the mean service time here is 1. The (infinite-server) offered load is given in (2.2). Since we have a sinusoidal arrival-rate function, we can apply Eick et al. (1993b) to give an explicit formula for the offered-load m_t^∞, i.e., the mean number of busy servers in the associated infinite-server system. Since the service-time distribution is exponential, we can apply formula (15) of Eick et al. (1993b). For

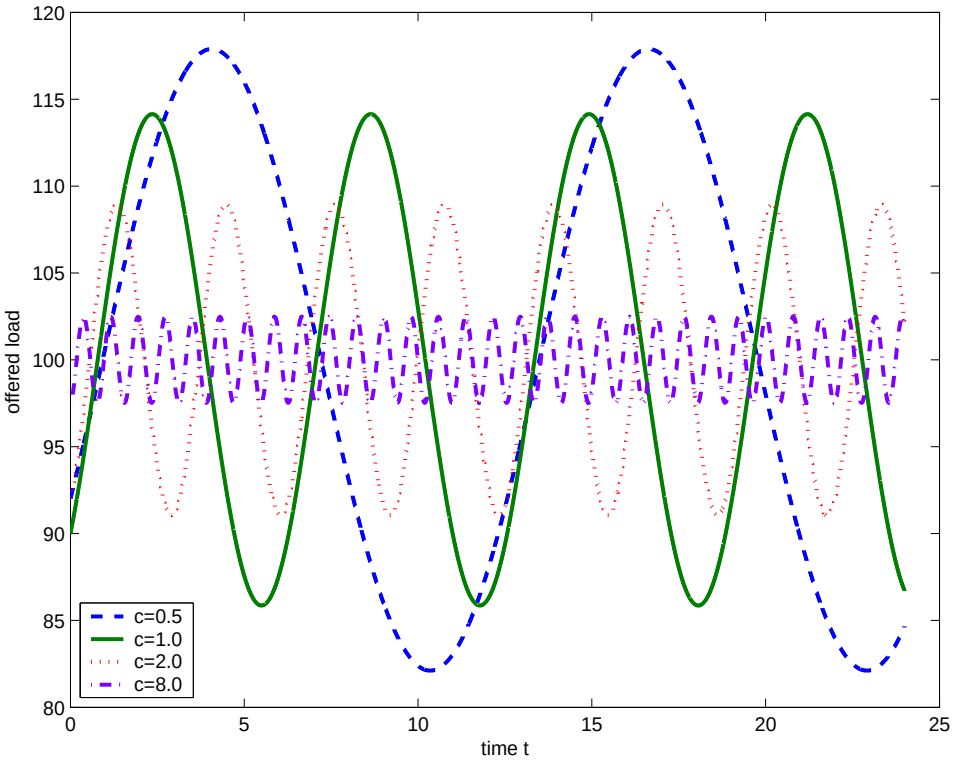
the sinusoidal arrival-rate function in (5.1), the offered load is

$$m_t^\infty = a + \frac{b}{1+c^2}[\sin(ct) - c \cdot \cos(ct)] = 100 + 10[\sin(t) - \cos(t)] \ . \tag{5.2}$$

The second formula in (5.2) is based on the specific parameters: $a = 100$, $b = 20$ and $c = 1$.

In order to put our model into perspective, in Figure 2 we plot the offered load m_t^∞ in (5.2) for the sinusoidal arrival-rate function in (5.1) for the parameters $a = 100$ and $b = 20$, as in our example, but with four different values of the time-scaling parameter c: 0.5, 1, 2 and 8. The offered load coincides with the mean number of busy servers in the $M_t/M/\infty$ model. Note that the offered load m_t^∞ is also a periodic function with the same period $2\pi/c$ as the arrival-rate function $\lambda(t)$. The length of the period decreases and the the size of the fluctuations decreases as c increases. As c increases, the modified offered load approaches the average value $a = 100$. It is important to understand the offered load, because it is a primary determinant of the required staffing, as we will see.

Figure 2: **The offered load m_t^∞ for the sinusoidal arrival-rate function in (5.1) with parameters $a = 100$, $b = 20$ and four possible values of c: 0.5, 1, 2 and 8.**



5.2. Application of the ISA

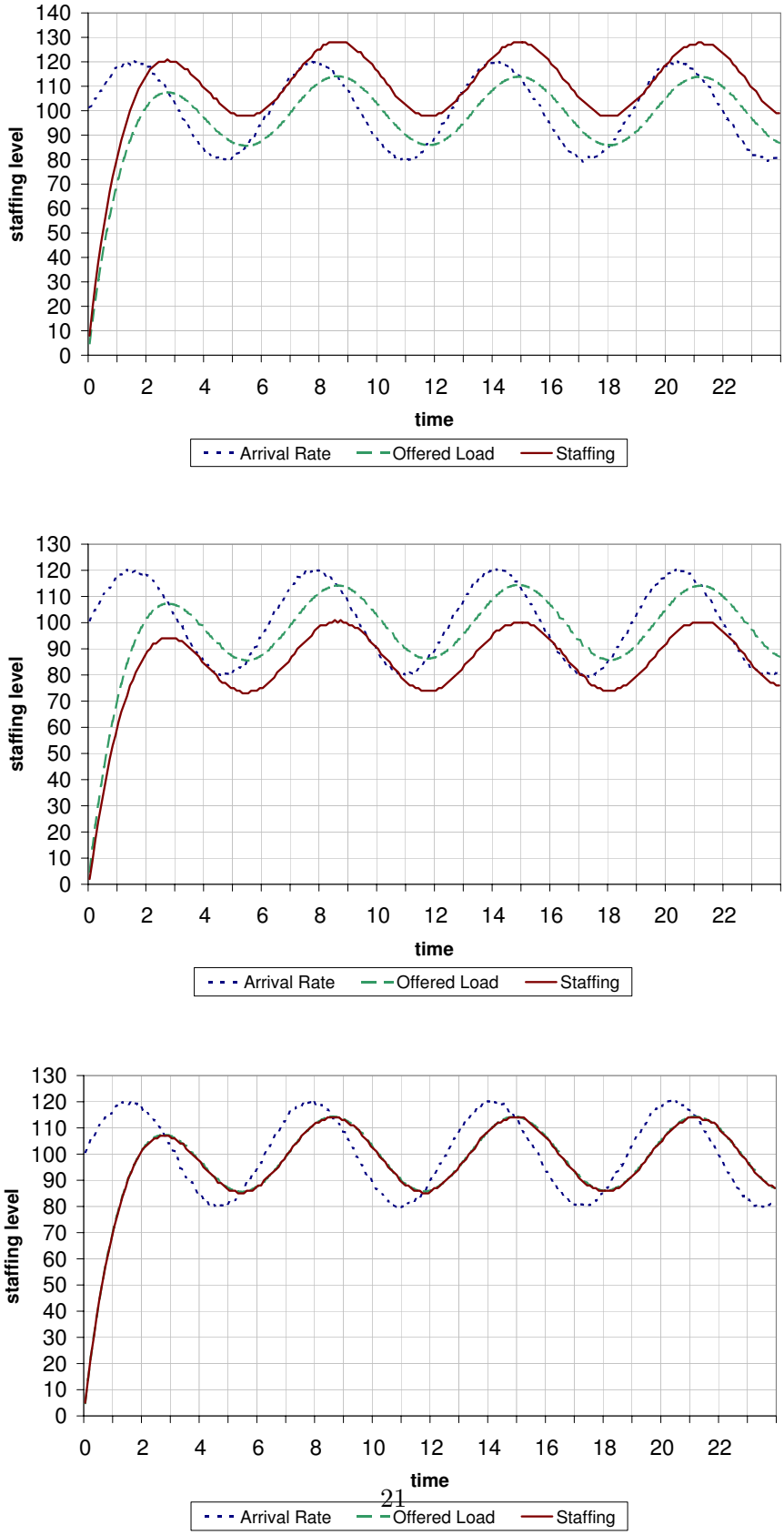
Our simulation-based iterated-staffing algorithm ISA generates staffing functions, for any given target delay probability α. In Figure 3 we present three graphs, showing the generated staffing functions for three regimes of operation: *Quality-Driven* (QD) - target $\alpha = 0.1$, *Efficiency-Driven* (ED) - target $\alpha = 0.9$, and *Quality-and-Efficiency-Driven* (QED) - target $\alpha = 0.5$.

In each graph, we plot three curves: the arrival rate $\lambda(t)$ (dotted), the offered load m_t (dashed) and the staffing function s_t (solid). Since the mean service time is 1 hour (and we are using ours as our time unit), the arrival rate $\lambda(t)$ here coincides with the PSA offered load $m_t^{PSA} \equiv \lambda(t)E[S]$, so the arrival rate and offered load are directly comparable. Note that the peak offered load lags behind (occurs later than) the peak arrival rate, but the ISA staffing follows the offered load. The ISA staffing levels in Figure 3 thus strongly support the square-root-staffing formula with the modified-offered-load (MOL) approximation.

Note that we start our system empty. This allows us to observe the behavior of the transient stage. In particular, there is a ramp-up at the left side of the plot. Our methods respond appropriately to that ramp-up. That is consistent with Section 7 of Jennings et al. (1996).

Note that the staffing level decreases as the target delay probability increases. Also note that, in the QED regime ($\alpha = 0.5$), the staffing function dictated by ISA falls right on top of the offered load: In that QED case, it would have sufficed to simply let $s_t = m_t^{\infty}$. The ISA-staffing s_t^{ISA} fell on top of the offered load m_t^{∞} in the QED regime, in particular when $\alpha = 0.5$, in all our experiments. That itself is quite stunning.

Figure 3: **Staffing for the time-varying Erlang-A example: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.9$ (ED), (3) $\alpha = 0.5$ (QED)**



5.3. Time-Stable Performance

We now show that ISA achieves time-stable performance. In Figure 4 we show the actual probability of delay obtained by applying our algorithm with target α for nine different values of the delay-probability target: $\alpha = 0.1, 0.2, \ldots, 0.9$. These delay probabilities are estimated by performing multiple (5000) independent replications with the final staffing function determined by our algorithm. Under the staffing levels produced by our algorithm, **the delay probabilities are remarkably accurate and stable**; the observed delay probabilities fluctuate around the target in each case.

In addition to stabilizing the delay probability, other performance measures (e.g. utilization and tail probabilities) are found to be quite stable as well. Precise explanations and definitions of the performance measures are given in Section 4.3. In Figures 5 are 6 are summary results graphs for all 9 target α's. These two performance measures increase as α increases, so we see the 9 cases starting with $\alpha = 0.1$ at the bottom and increasing to the case $\alpha = 0.9$ at the top.

However, as the target delay probability increases toward heavy loading, the abandonment probability becomes much less time-stable, as shown in Figure 7. We discuss this phenomenon further in §10 below.

Figure 4: **Delay probabilities for the sinusoidal example with nine delay-probability targets α, ranging from 0.1 to 0.9.**

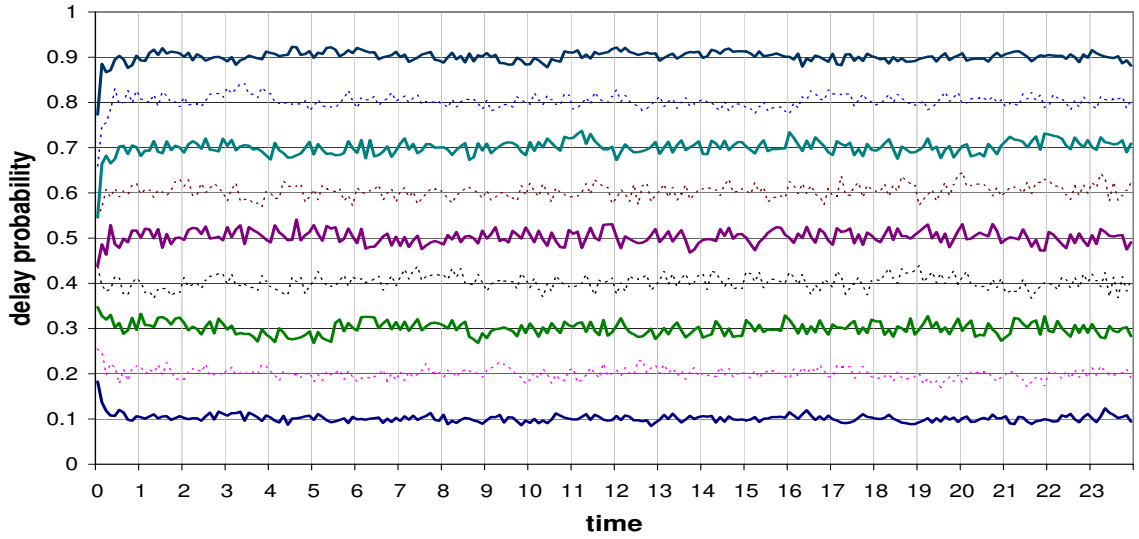
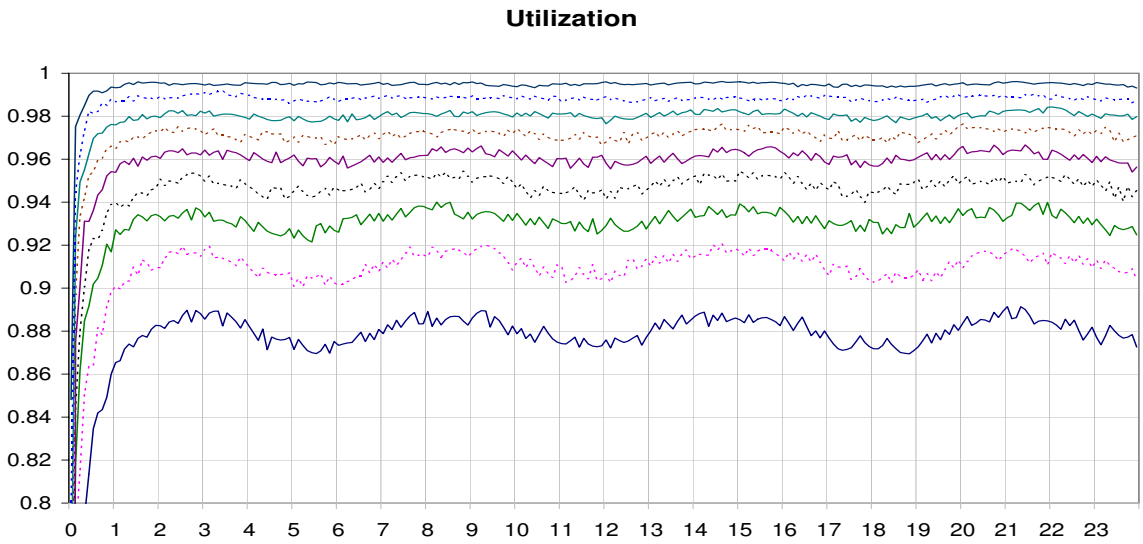


Figure 5: **Utilization summary for the time-varying Erlang-A example**



Other measures of congestion such as average waiting time and average queue length were also found to be relatively stable, but not perfectly so; e.g., see Figure 8.

Figure 6: **Tail probability summary for the time-varying Erlang-A example**

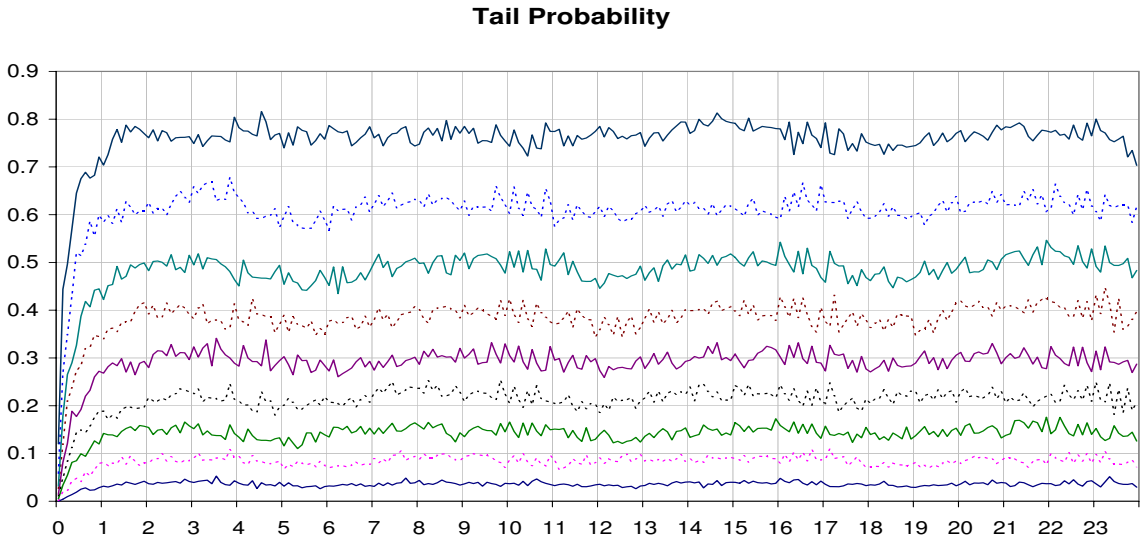


Figure 7: **Abandonment probabilities for the same sinusoidal example with the same nine delay-probability targets.**

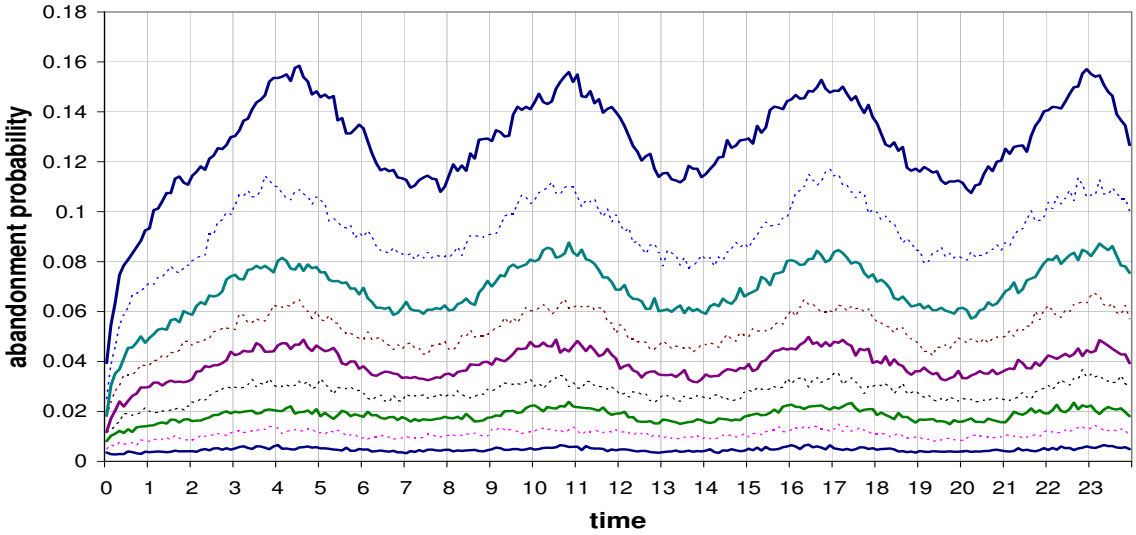
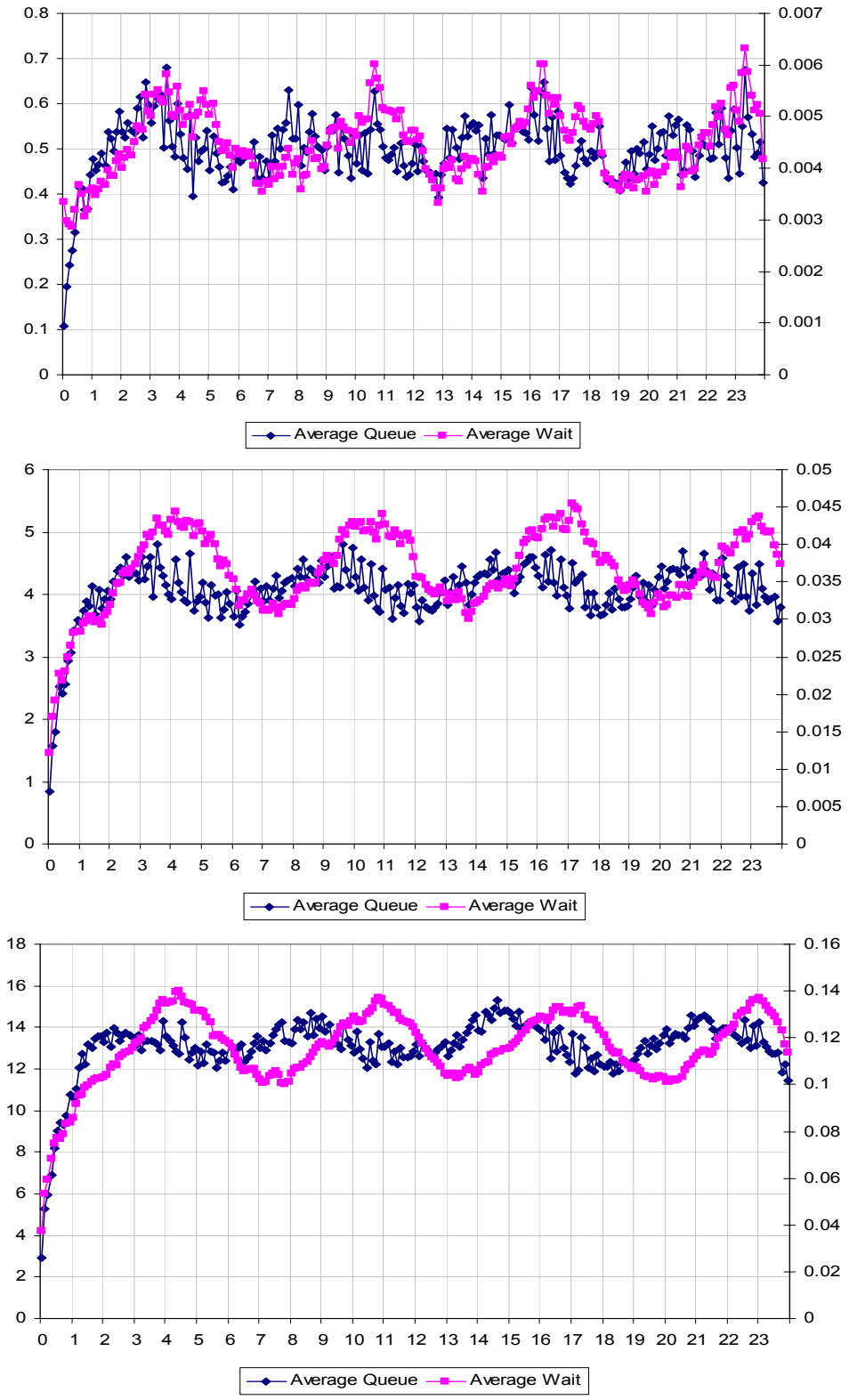


Figure 8: **Congestion measures in the time-varying Erlang-A example in the three regimes: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.5$ (QED), (3) $\alpha = 0.9$ (ED)**



5.4. Validating the Square-Root-Staffing Formula

We now validate the square-root-staffing formula in (2.7). For that purpose, we define an **implied empirical service quality**: A function $\{\beta_t^{ISA} : 0 \le t \le T\}$ is defined by setting

$$\beta_t^{ISA} \equiv \frac{s_t^{ISA} - m_t^\infty}{\sqrt{m_t^\infty}}, \quad 0 \le t \le T \ , \tag{5.3}$$

where m_t^∞ is again the offered load in (2.2) and (5.2). while s_t^{ISA} is the staffing function obtained by the ISA algorithm. Since s_t^{ISA} is obtained from the ISA algorithm, the function β_t^{ISA} is itself obtained from the ISA algorithm. It thus becomes interesting to see if **the implied service quality is approximately constant as a function of time**, because that would imply that (5.3) approximately coincides with the square-root-staffing formula (2.7). And, indeed, it is, as shown in Figure 9. Again we consider 9 values of α ranging from 0.1 to 0.9 in steps of 0.1. As α increases, the quality of service reflected by β_t^{ISA} decreases. But the main point is that the empirical service quality β_t as a function of t is approximately constant as a function of t for each α over the full range from 0.1 to 0.9.

Figure 9: **The implied empirical quality of service β_t^{ISA} for the sinusoidal example, decreasing as α increases through the values** 0.1, 0.2, . . . , 0.9.

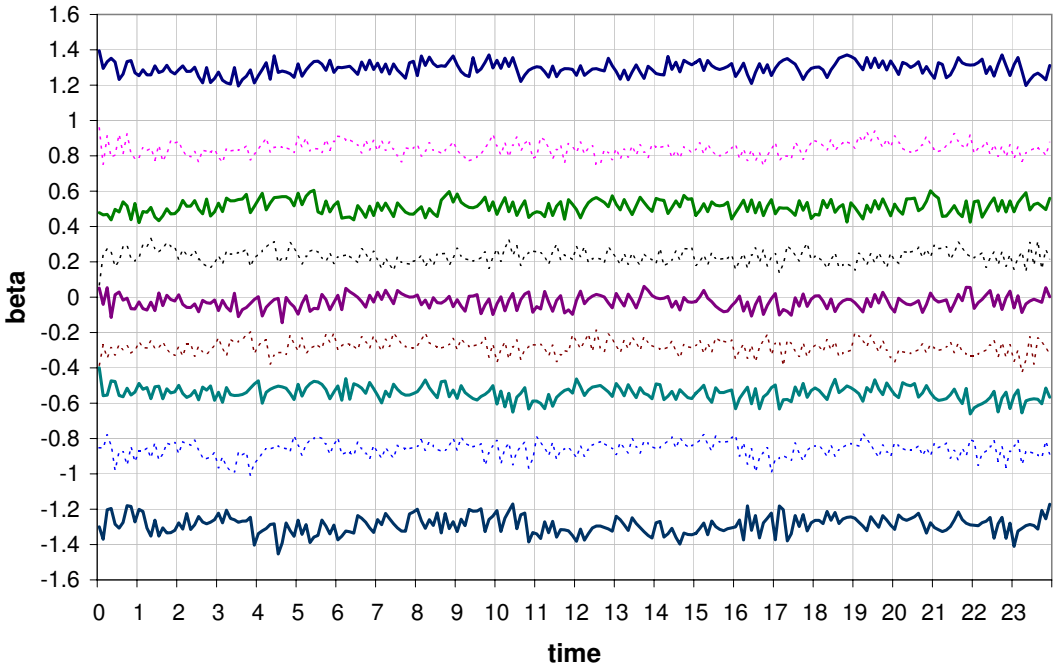


Figure 9 is extremely important because it validates the square-root-staffing formula for this example. First, Figure 4 shows that ISA is able to produce the target delay probability α for a wide range of α. Then Figure 9 shows that, when this is done, the square-root-staffing formula holds empirically. In other words, we have shown that we could have staffed directly by the square-root-staffing formula instead of by the ISA.

5.5. Relating β to α

However, one issues remains: In order to staff directly by the square-root staffing formula, *we need to be able to relate the quality of service β to the target delay probability α*. Indeed, we want a function mapping α into β. We propose a simple answer: For the time-varying Erlang-A model, **use the associated stationary Erlang-A model**, i.e., the $M/M/s+M$ model. That is tantamount to applying the **modified-offered-load approximation** to the $M/M/s+M$ model. Previously the MOL approximation has been applied only to the pure-loss and pure-delay models (without customer abandonments).

Moreover, we suggest using simple formulas obtained from the **many-server heavy-traffic limit** for the Erlang-A model in Garnett et al. (2002). The **Garnett-Mandelbaum-Reiman function, for brevity here referred to as the Garnett function** mapping β into α is

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \qquad -\infty < \beta < \infty; \tag{5.4}$$

where $\hat{\beta} = \beta\sqrt{\theta/\mu}$, with μ the individual service rate and θ the individual abandonment rate (both here set equal to 1 now) and $h(x) = \phi(x)/(1-\Phi(x))$ is the *hazard rate* of the standard normal distribution, with ϕ being the *probability density function* (pdf) and Φ the cdf. Of course, we want a function mapping α into β. Thus, we use the **inverse of the Garnett function**, which is well defined.

We now look at additional simulation output, aimed at establishing the validity of this stationary-model approach of relating α and β. First, we compare the empirical distribution of the customer waiting times to the theoretical distribution of those waiting times in the stationary Erlang-A model. Specifically, in Figure 10 we plot the *empirical conditional waiting time pdf*, given wait, i.e. the distribution of the waiting time for those who were in fact delayed, during the entire time-horizon. In doing so, we are looking at all the waiting times experienced across the day. As before, we obtain statistically precise estimates by averaging over a large number of independent replications (here again 5000). In this case, the empirical conditional

distribution is based on statistics gathered from the time of reaching steady until the end of the horizon.

In Figure 10 we compare the empirical conditional waiting-time distribution to many-server heavy-traffic approximations for the conditional waiting-time distribution in the **stationary** $M/M/s+M$ **queue**, drawing on Garnett et al. (2002). Note that the approximation for the conditional waiting-time distribution in the stationary queues matches the performance of our time-varying model remarkably well.

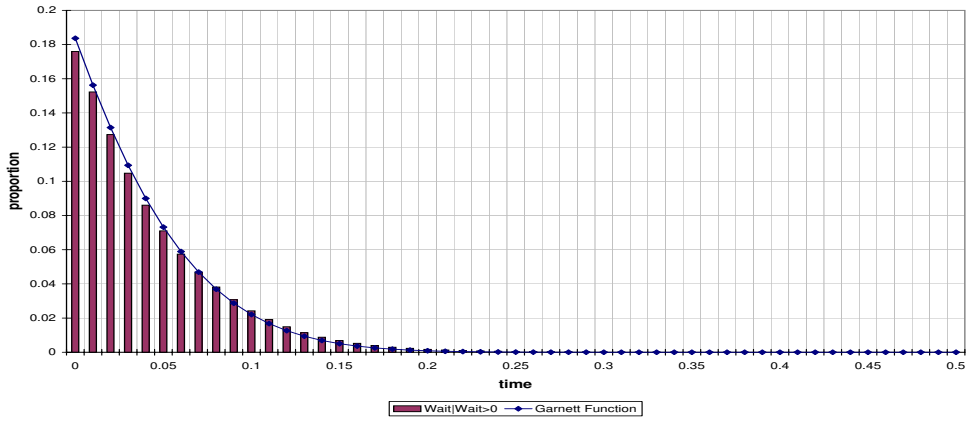
The significance of what we are seeing deserves some additional discussion. First, in the way of background, we note that the stationary waiting-time distribution in the $M/M/s+M$ model, for parameters λ, s and θ with $\mu = 1$ and $s \approx (\lambda/mu) + \beta\sqrt{\lambda/\mu}$ tends to depend only on the parameters β and θ, provided that s and λ are not small and $|\beta|$ is not large, as theoretically deduced by the many-server heavy-traffic limit in Garnett et al. (2002). For $\theta = 1$ as we are considering, then, the single key parameter is β. In other words, as long as we keep the parameters β and θ fixed, the stationary distribution tends to be approximately independent of s or λ (with the relation between them depending on β).

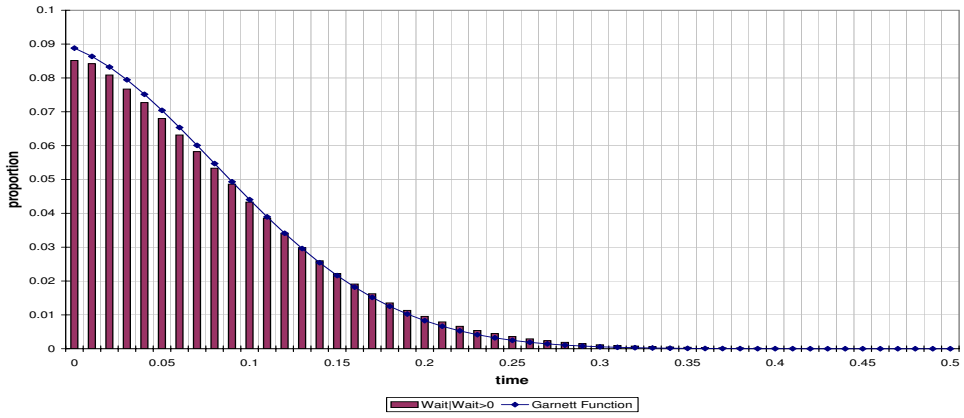
Second, the time-dependent distribution in the $M_t/M/s_t+M$ model with $s_t \approx m_t + \beta\sqrt{m_t}$ and β held fixed, tends to have the same distribution at each t as the stationary model discussed above, with the same β. That is being supported by Figure 10. Indeed, the beautiful plots we see in Figure 10 remain valid if we only sample at selected subsets of the times, e.g., only when the arrival rate is near its maximum or only when the arrival rate is near its minimum.

We next relate the empirical (α, β) pairs to the Garnett function in (5.4). We define the empirical values $\bar{\alpha}$ and $\bar{\beta}$ as simply the time-averages of the observed (time-stable) values displayed in the plots in Figures 4 and 9. In Figure 11, we plot the pairs of $(\bar{\alpha}_i, \bar{\beta}_i)$ alongside the Garnett function. Needless to say, the agreement is phenomenal!

As a consequence, we see that we can use the inverse of the Garnett function in (5.4) and in Figure 11 to calculate the appropriate quality-of-service parameter β to use in the square-root-staffing formula (2.7) for any specified target delay probability α.

Figure 10: **The empirical conditional waiting time distribution, given positive wait, for the $M_t/M/s_t + M$ example with three delay-probability targets: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.5$ (QED), (3) $\alpha = 0.9$ (ED).**





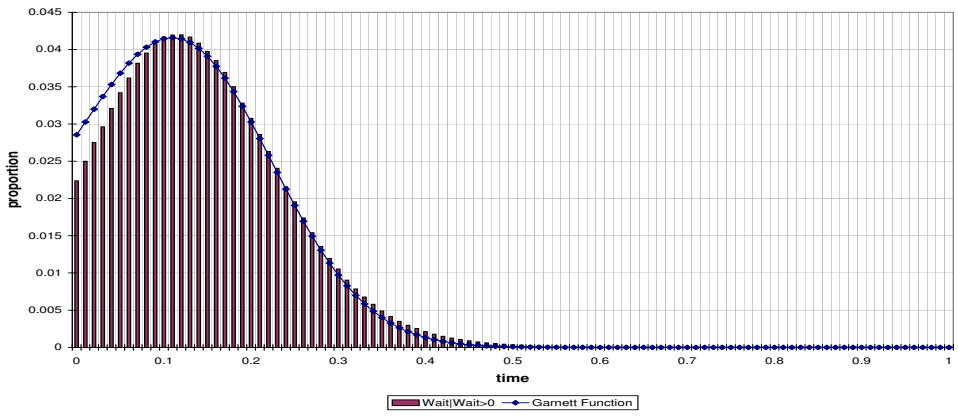
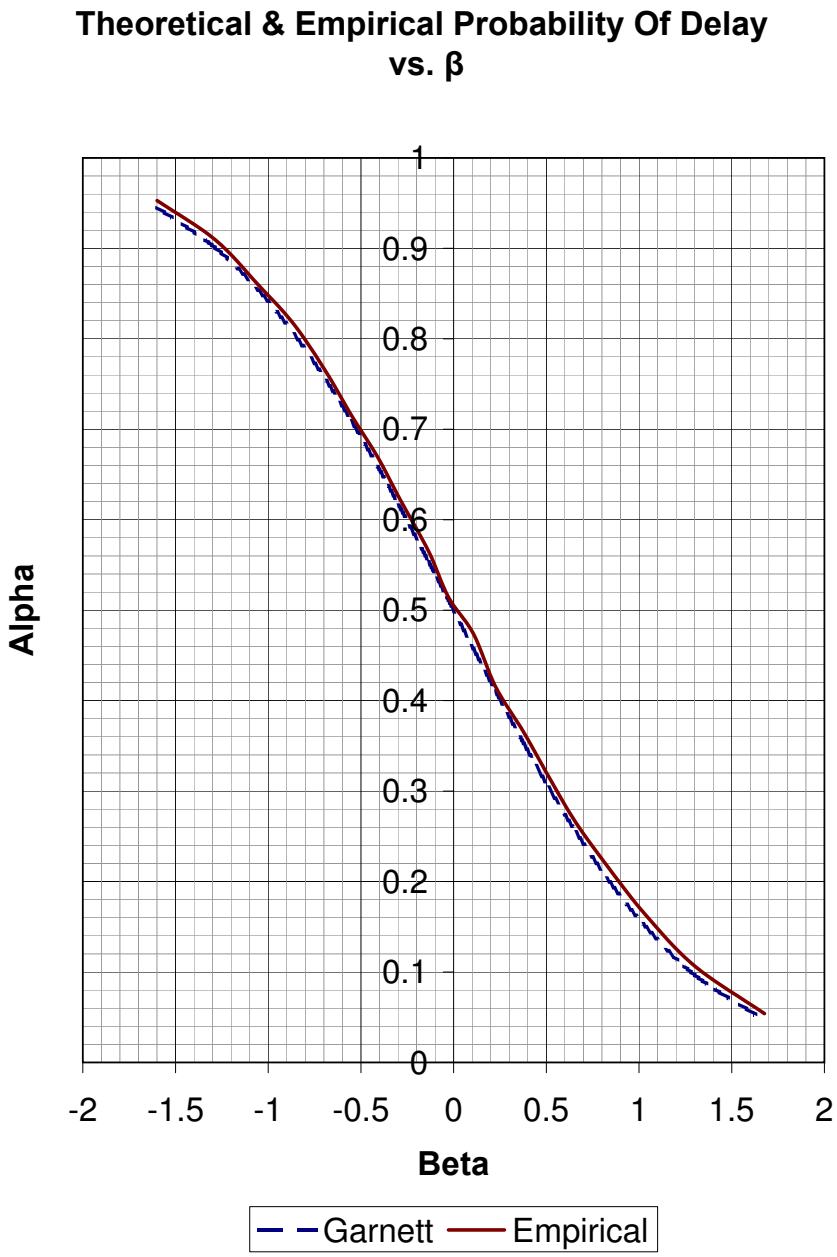


Figure 11: **A comparison of the empirical relation between α and β with the Garnett function for the time-varying Erlang-A example**



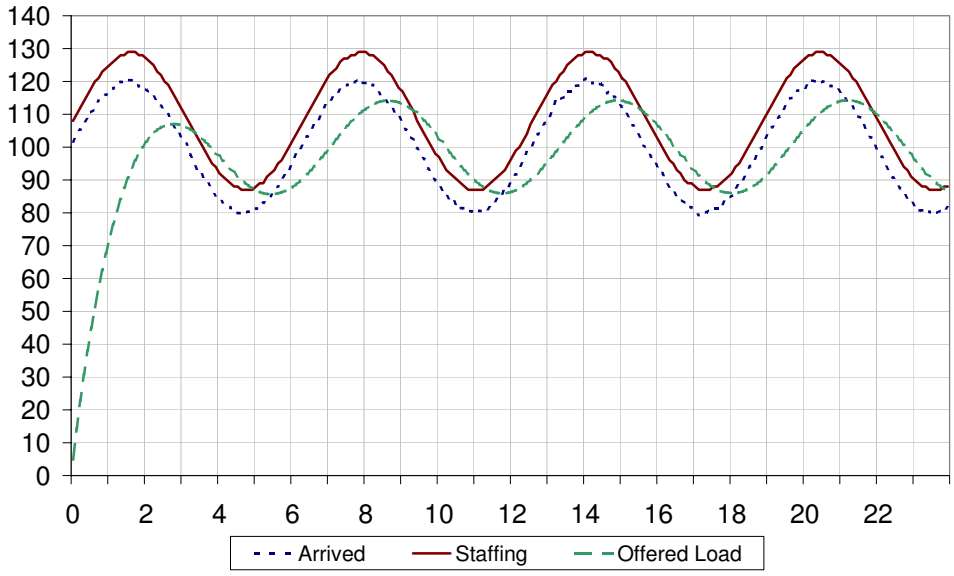
5.6. Comparison to PSA and SSA

Jennings et al. (1996) observed that their new infinite-server approximation and modified-offered-load approximation both performed much better than classical alternatives, namely, the pointwise stationary approximation (PSA) and the simple stationary approximation (SSA); SSA uses the stationary model with the overall long-run average arrival rate.

We now show that the same is true here for the time-varying Erlang-A model (using the same parameters as before: $\mu = \theta = 1$ and the sinusoidal arrival-rate function in (5.1) with $a = 100$, $b = 20$ and $c = 1$). First we plot the arrival rates and staffing levels for PSA and SSA in Figure 5.6. Then we plot the delay probabilities and the mean queue lengths and mean waiting times for the two methods in Figures 5.6 and 5.6. The specific target delay probability here is $\alpha = 0.2$. That can be confirmed by looking at SSA. The arrival rate $\lambda(t)$ has average 100. For $s = 109$, the steady-state delay probability in the $M/M/s+M$ model with $\mu = \theta = 1$ is 0.196; while for $s = 108$, the steady-state delay probability is 0.225. Incidentally, the steady-state abandonment probabilities in those two cases are 0.0104 and 0.0124, respectively. In contrast, in the nonstationary environment we see that the time-dependent delay probablity is as much as 0.75, with an average over 0.35. From the customer's perspective, the inconsistency of performance may be as bad as the high congestion itself.

The delay probability plots are quite similar for PSA and SSA, but note that the peaks and troughs do not occur at the same places. From Figure 5.6, we see that PSA tends to understaff most when going down, at a point just below the average (100), e.g., at time 10.5, while SSA tends to understaff most when MOL is near its peak, e.g., at time 8.5.

Figure 12: **Staffing levels for (1) the pointwise-stationary approximation (PSA) and (2) the simple-stationary approximation (SSA) for the time-varying Erlang-A example**



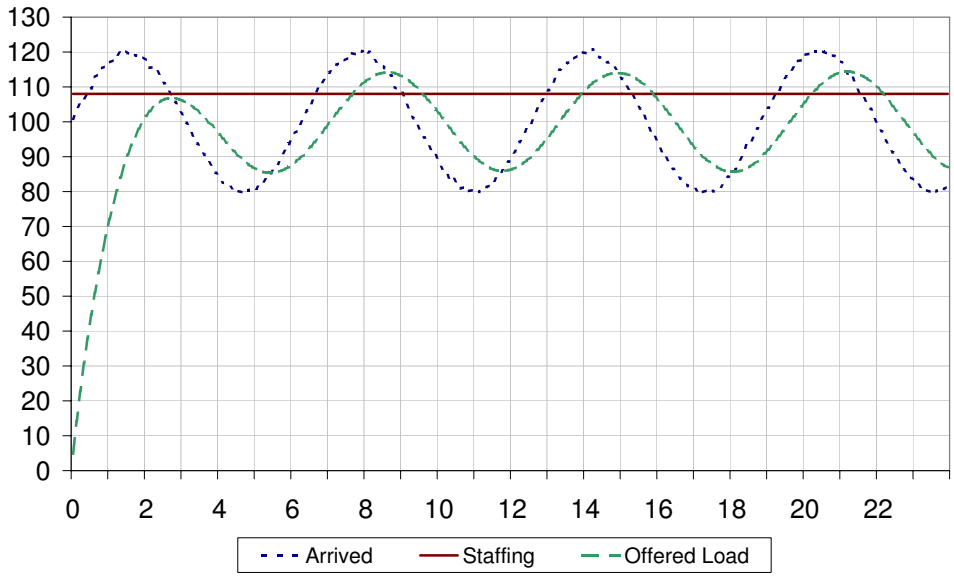
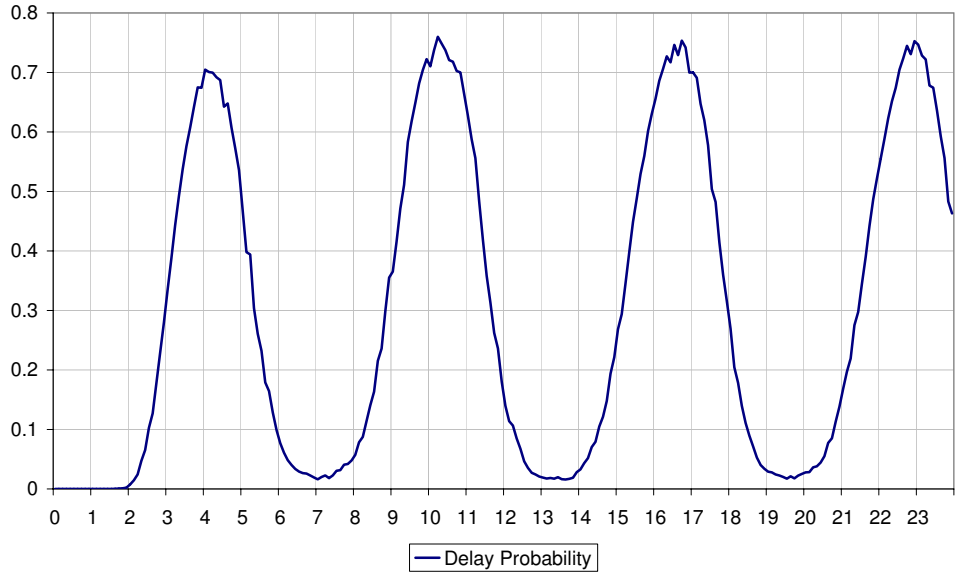


Figure 13: **Delay probabilities for (1) the pointwise-stationary approximation (PSA) and (2) the simple-stationary approximation (SSA) for the time-varying Erlang-A example**



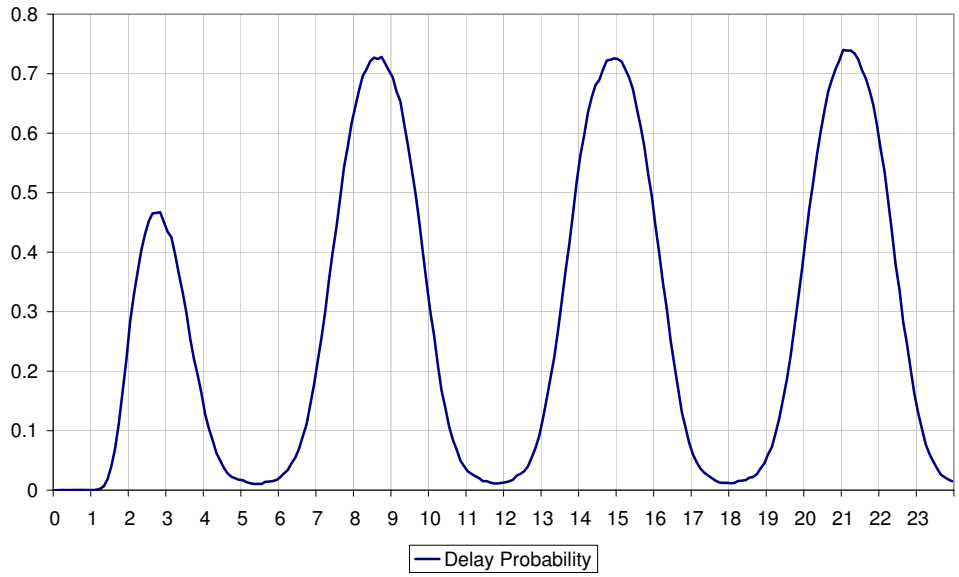
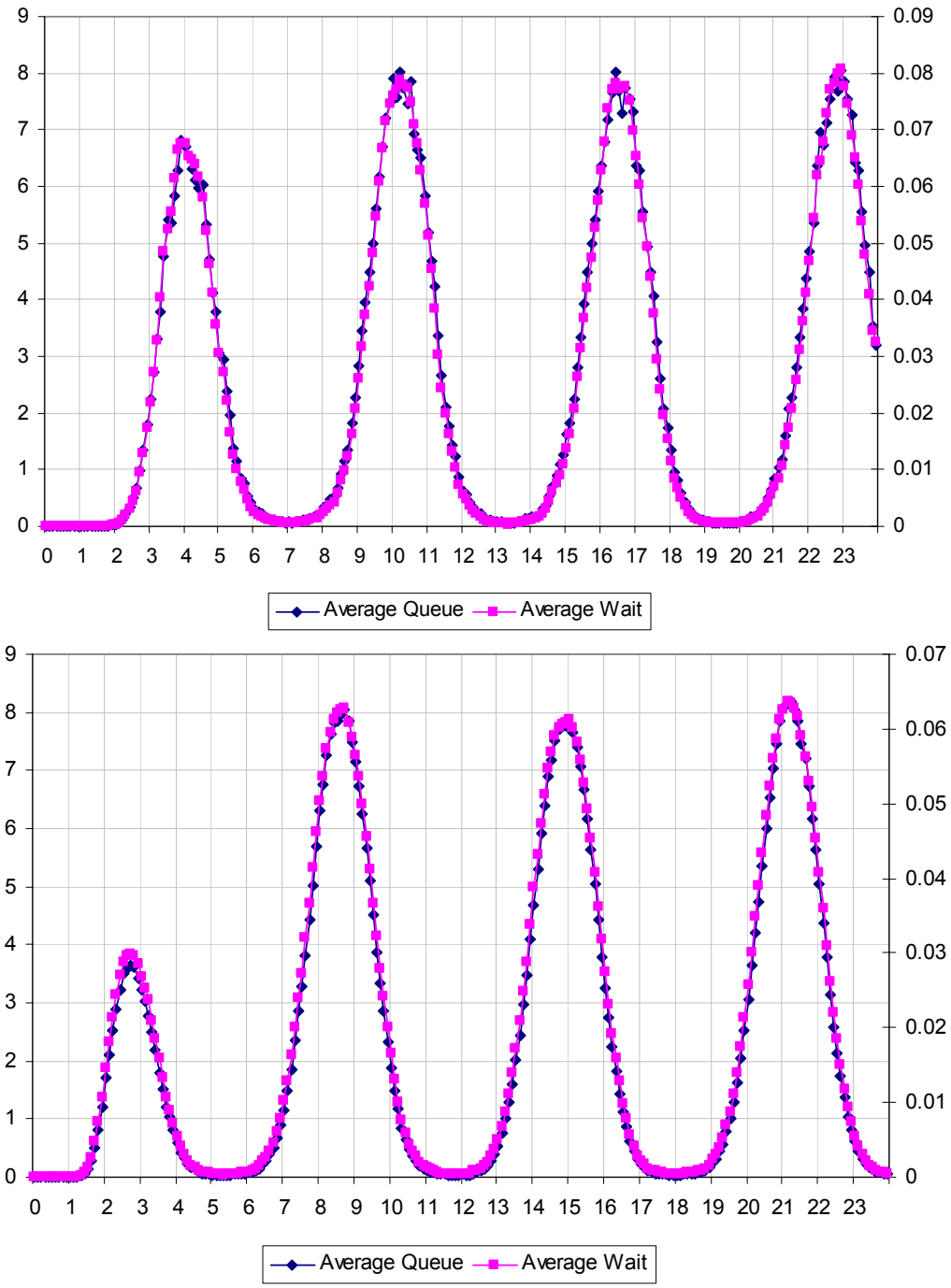


Figure 14: **Mean queue lengths and mean waiting times for (1) the pointwise-stationary approximation (PSA) and (2) the simple-stationary approximation (SSA) for the time-varying Erlang-A example**



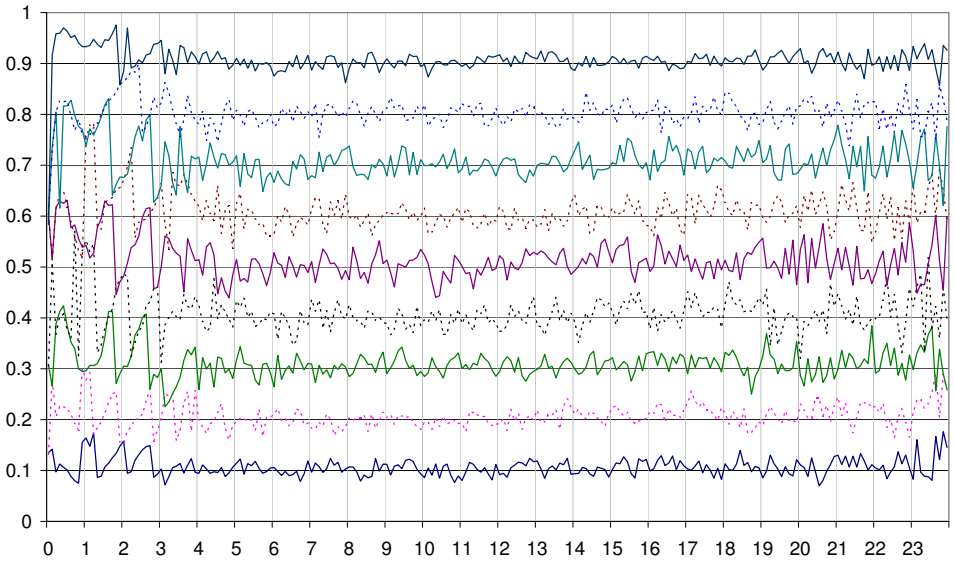
6. The Realistic Example Related to Figure 1

In this section we consider the practical case that was first described in Figure 1. This example is more realistic than the previous example, not only because we use an actual arrival-rate function, but also because we use more realistic (shorter) service times. Specifically, we decrease the mean service time from 1 hour to 6 minutes. That is achieved with our hourly time scale by letting $\mu = 10$. Corresponding to that, we let $\theta = 10$, so that we have $\theta = \mu$ as in Section 5. Results are shown below.

6.1. Time-Stable Delay Probabilities Again from ISA

We first plot the delay probabilities and implied empirical quality of service β_t^{ISA} for the nine target delay probabilities ranging from 0.1 to 0.9, just as before. They appear in Figures 15 and 16. We see that we again achieve time-stable performance with ISA, and again it agrees with MOL.

Figure 15: **Delay probability summary for the realistic example**



With such short service times, we might think that that this should be an easy problem, for which simple PSA would also work well. Indeed, when we look at the staffing for three values of α in Figure 17, we do not see much difference, but there actually is a difference. Even though the service times are indeed short here, the arrival-rate function is changing rapidly at some times, especially in hours $4 - 6$. For this example, Figure 18 shows that simple PSA performs significantly worse than ISA.

Figure 16: **Implied service quality β summary for the realistic example**

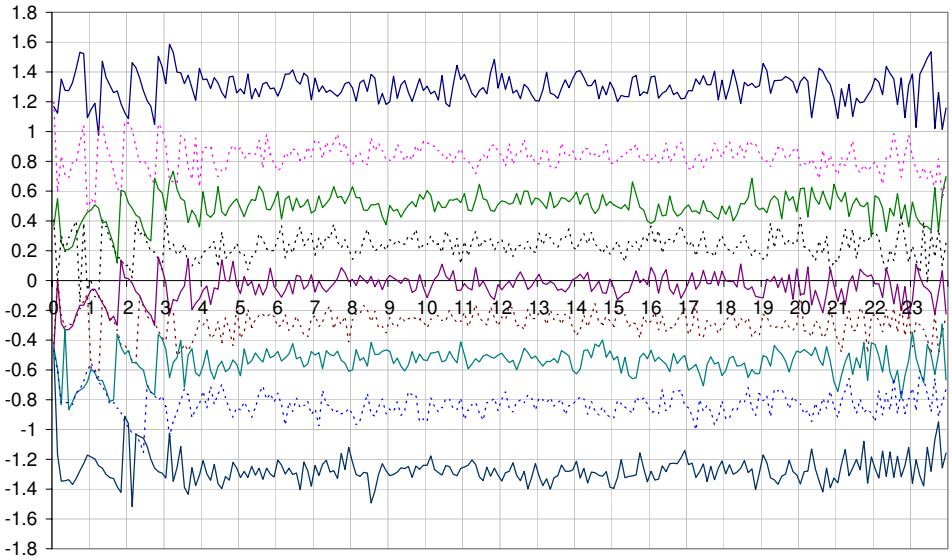
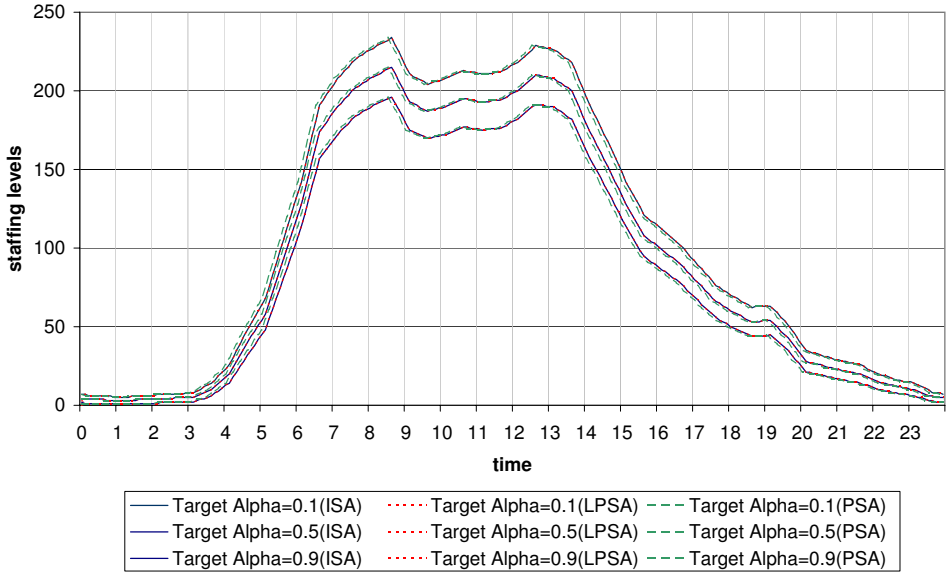


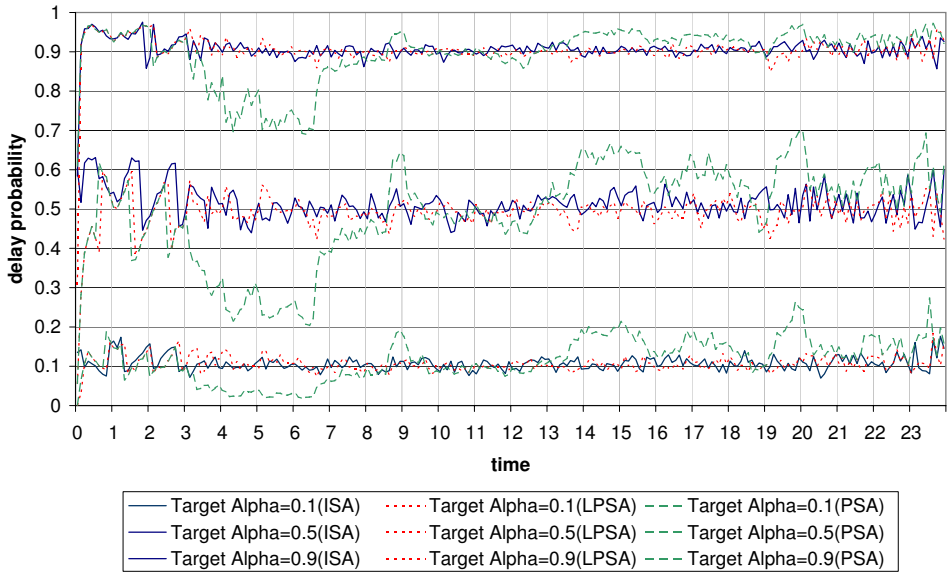
Figure 17: **A comparison of staffing levels based on ISA, PSA and lagged PSA for the realistic example, for three delay-probability targets: 0.1, 0.5 and 0.9.**



6.2. Lagged PSA

As before, we find that ISA produces essentially the same results as MOL. Moreover, the dominant effect in MOL is captured by the time lag in (2.4); i.e., here it suffices to use *lagged PSA*, with approximate offered load $\lambda(t - E[S_e])E[S]$. Since the service-time distribution is

Figure 18: **A comparison of ISA, PSA and lagged PSA for the same three delay-probability targets.**



exponential, S_e and S have a common exponential distribution, and the lagged-PSA offered load is just $\lambda(t - E[S])E[S]$. The good performance of lagged PSA is consistent with the various refinements proposed by Green et al. (2001). We show that simple PSA performs worse than ISA and lagged PSA by plotting the delay probabilities for these three staffing rules in Figure 18. The performance of simple PSA here is nowhere near as bad as it was in the challenging $M_t/M/s_t$ example in Jennings et al. (1996), and as it is for the example here in §5 (see §5.6), but there are clear departures from the performance targets in Figure 18. The PSA delay probabilities are significantly below the targets during the hours $4 - 6$ with rapidly increasing arrival rates. The differences among the corresponding staffing functions in Figure 17 look small, but those small differences can have a significant impact, because the arrival-rate function changes rapidly.

6.3. Deviations at the Ends of the Day

We also observe that ISA is not as successful as before, because the target delay probability is not achieved accurately at the beginning and at the end of the day. This phenomenon is even more evident for other performance measures, as shown in the plots below.

However, this bad behavior is quite clearly due to the extremely low arrival rates that prevail at the beginning and the end of the day. When the load is small, the addition or

Figure 19: **Abandon probability summary for the realistic example**

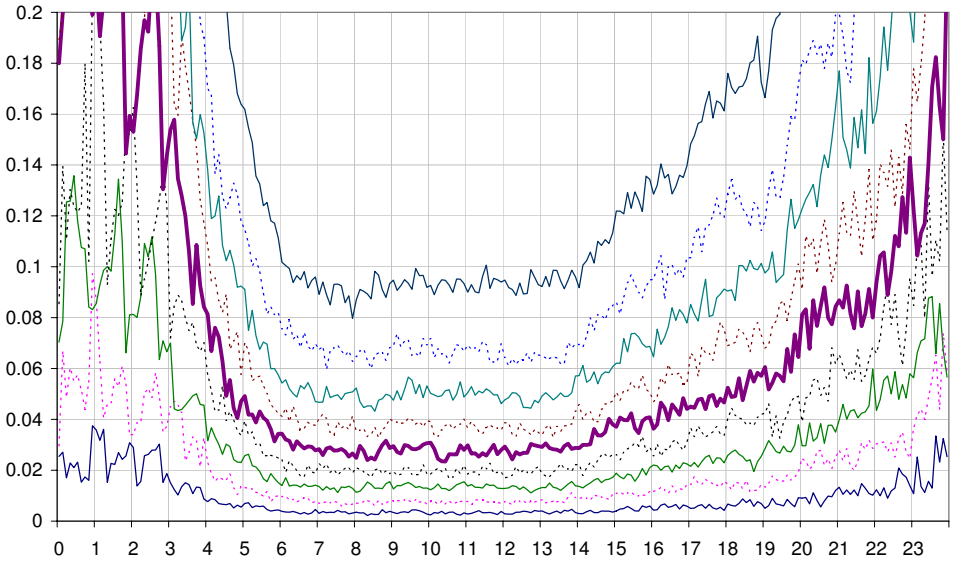
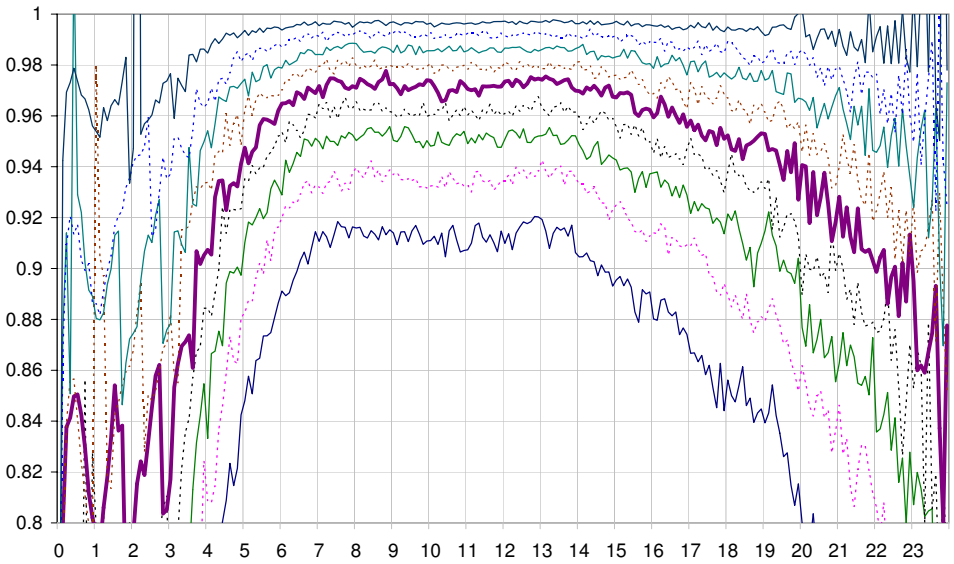


Figure 20: **Utilization summary for the realistic example**



removal of a single server while greatly affect the delay probability. On the positive side, note that there is a clear time-interval - from 5 to 18, in which performance measures are quite stable, and when operating under reasonable service quality (up to delay probability of 0.5), performance measures are varying in quite a small range.

In Figures 37-39 we further describe the performance of ISA in the three regions: QD

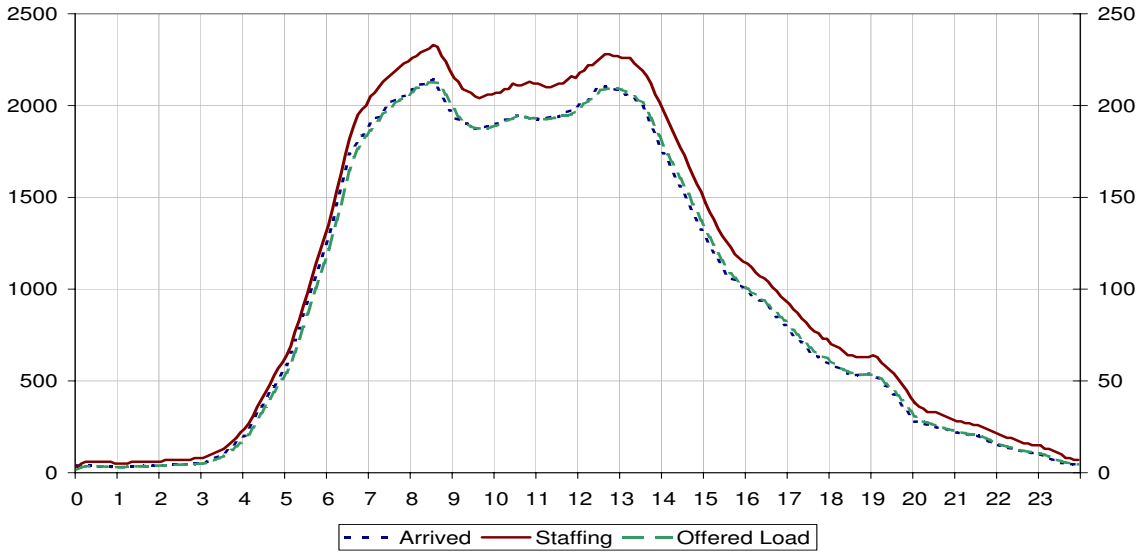
($\alpha = 0.1$), ED ($\alpha = 0.9$) and QED ($\alpha = 0.5$). There are several important observations to make here: First, note that in all cases the (infinite-server) offered load m_t seems to fall almost directly on top of the PSA offered load $m_t^{PSA} \equiv \lambda(t)/\mu$, suggesting that in this case that the MOL approximation essentially coincides with the elementary PSA approximation. But from Figure 18 we know that we need to incorporate the lag in PSA (to get lagged PSA) in order to get that good performance. But the square-root-staffing formula (2.7) will perform the same using either the infinite-server offered load or the lagged-PSA offered load. Moreover, the ISA performs the same as the square-root-staffing formula with either the infinite-server offered load or the lagged-PSA offered load.

Consequently, ISA does not differ much from lagged-PSA. However, for the time-varying Erlang-A model, staffing using lagged-PSA is actually not so routine. We need to apply the steady-state distribution of the $M/M/s + M$ model or a suitable approximation.

The three regimes of operation in Figures 37-39 are clearly revealed by the average waiting time: In the QD regime the average waiting time is relatively negligible; in the QED regime average waiting time is in seconds; and in the ED it is in minutes. Figure 39 shows, once again, that the staffing falls right on top of the offered load in the QED regime.

Finally, Figure 40 shows that the excellent matching between the Garnett function and the empirical results is preserved also in this example.

Figure 21: **The realistic example in the QD regime (target α=0.1): (1) staffing level, offered load and arrival function, (2) average queue length and average waiting time (in average service time)**



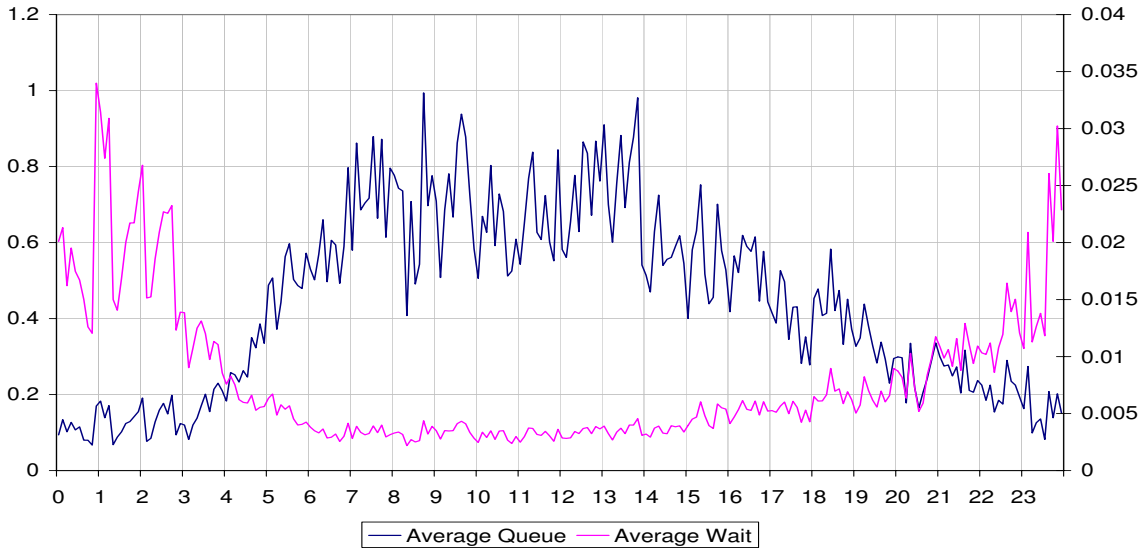
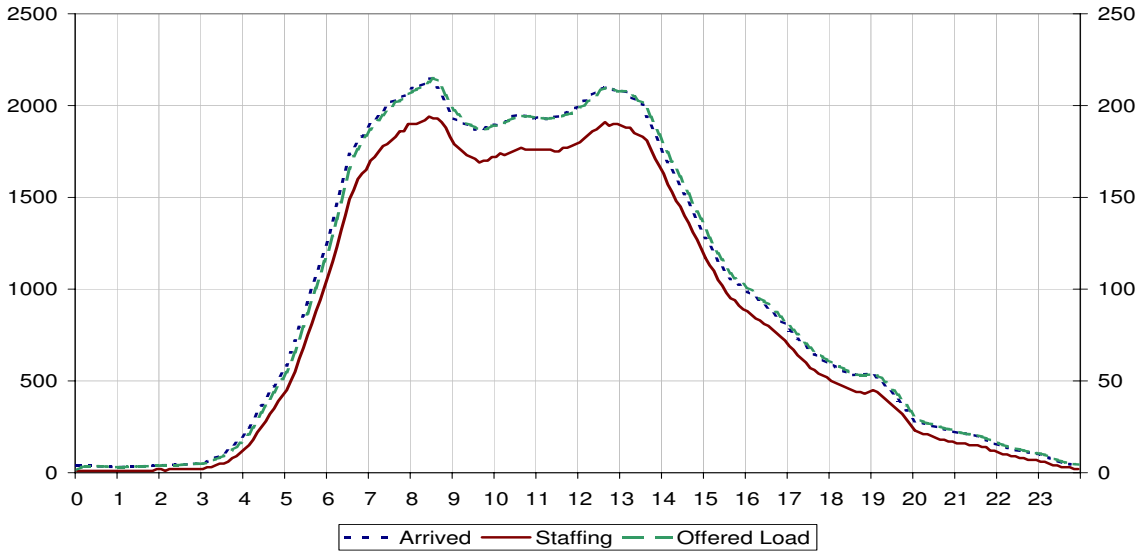


Figure 22: **The realistic example in the ED regime (target α=0.9): (1) staffing level, offered load and arrival function, (2) average queue length and average waiting time (in average service time)**



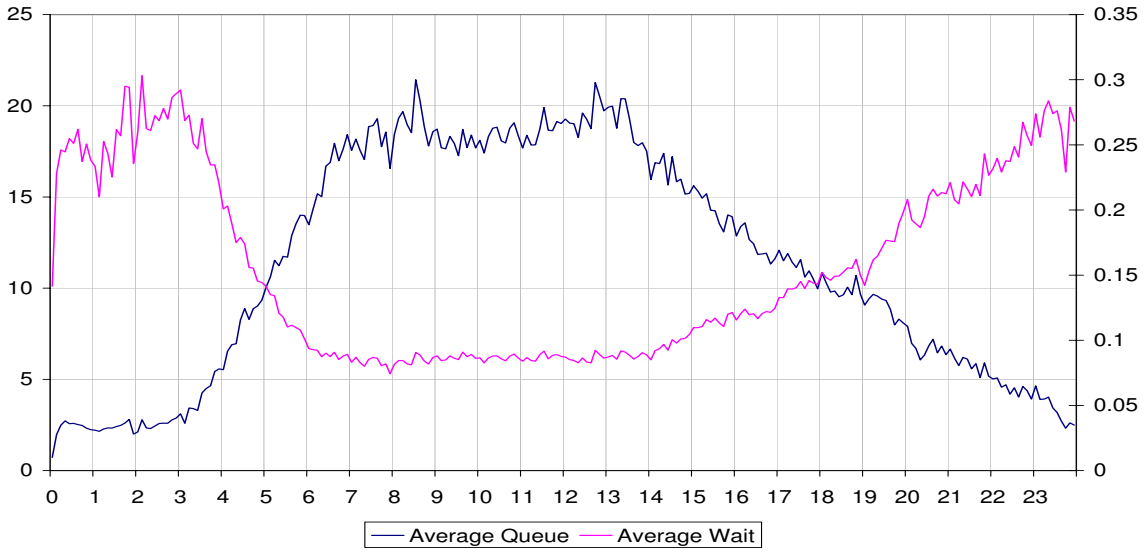
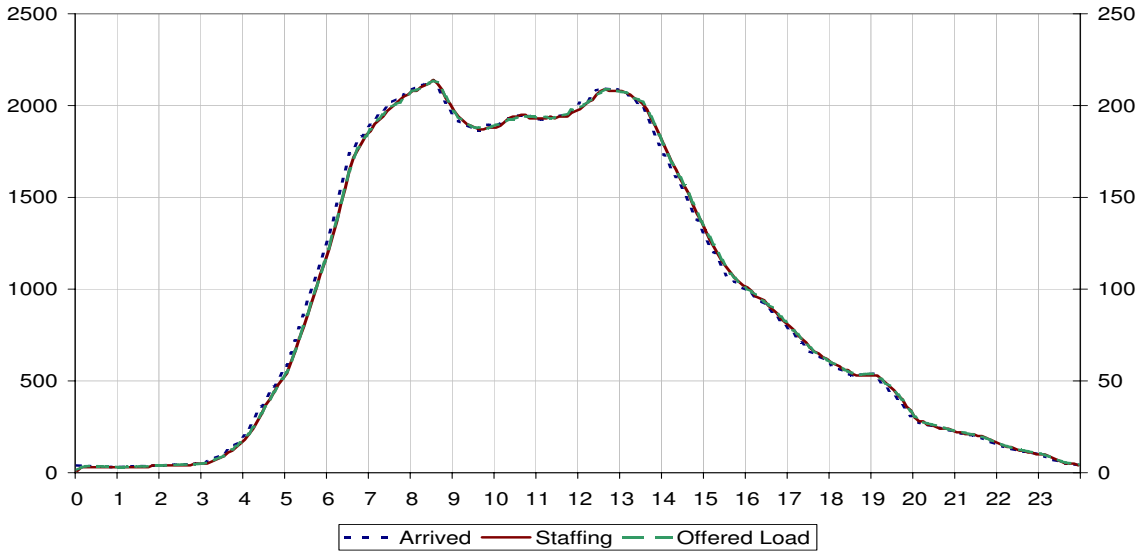


Figure 23: **The realistic example in the QED regime (target α=0.9): (1) staffing level, offered load and arrival function, (2) average queue length and average waiting time (in average service time)**



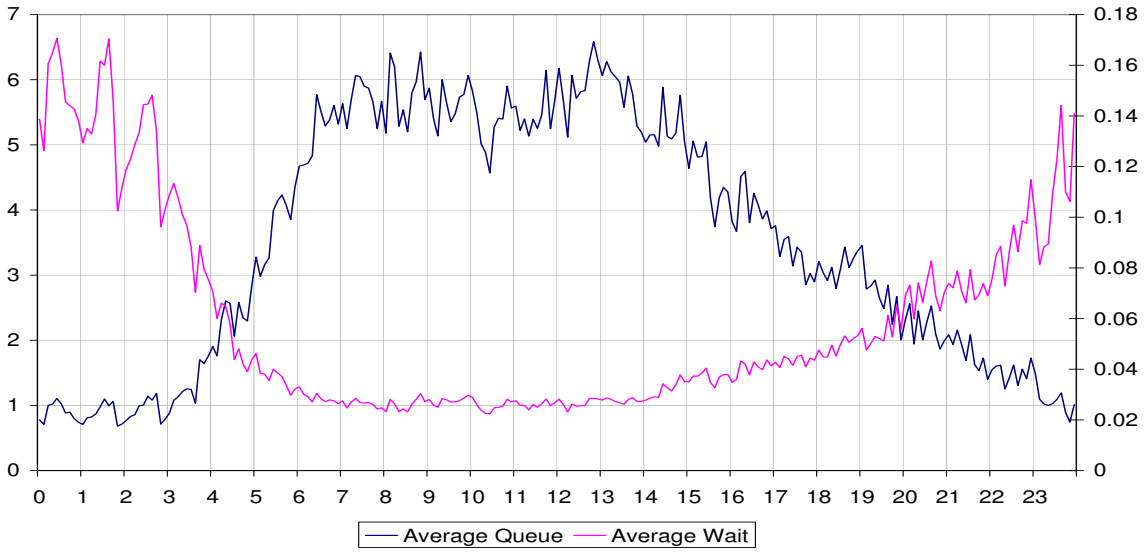
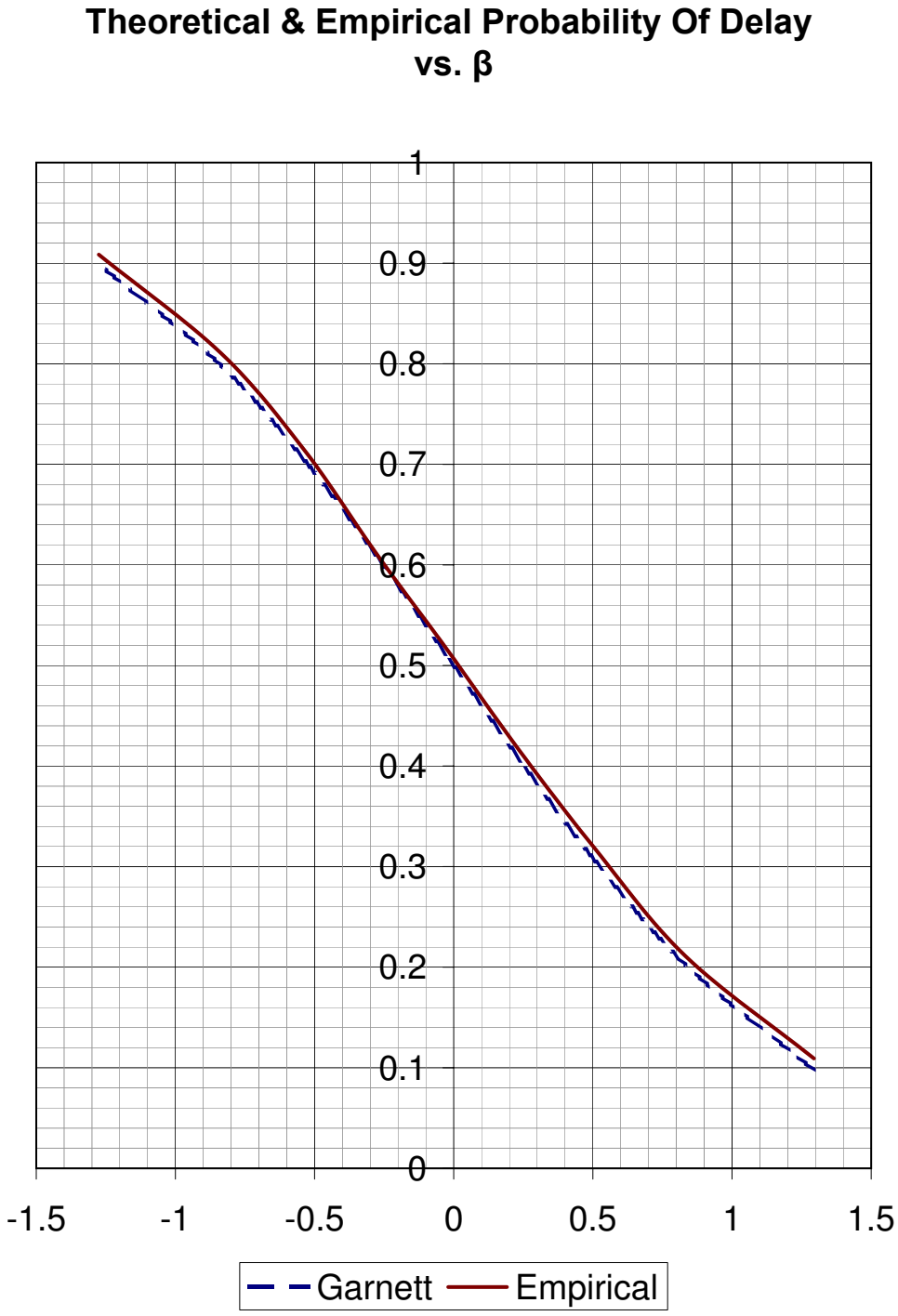


Figure 24: **Comparison of empirical results with the Garnett approximation for the realistic example**

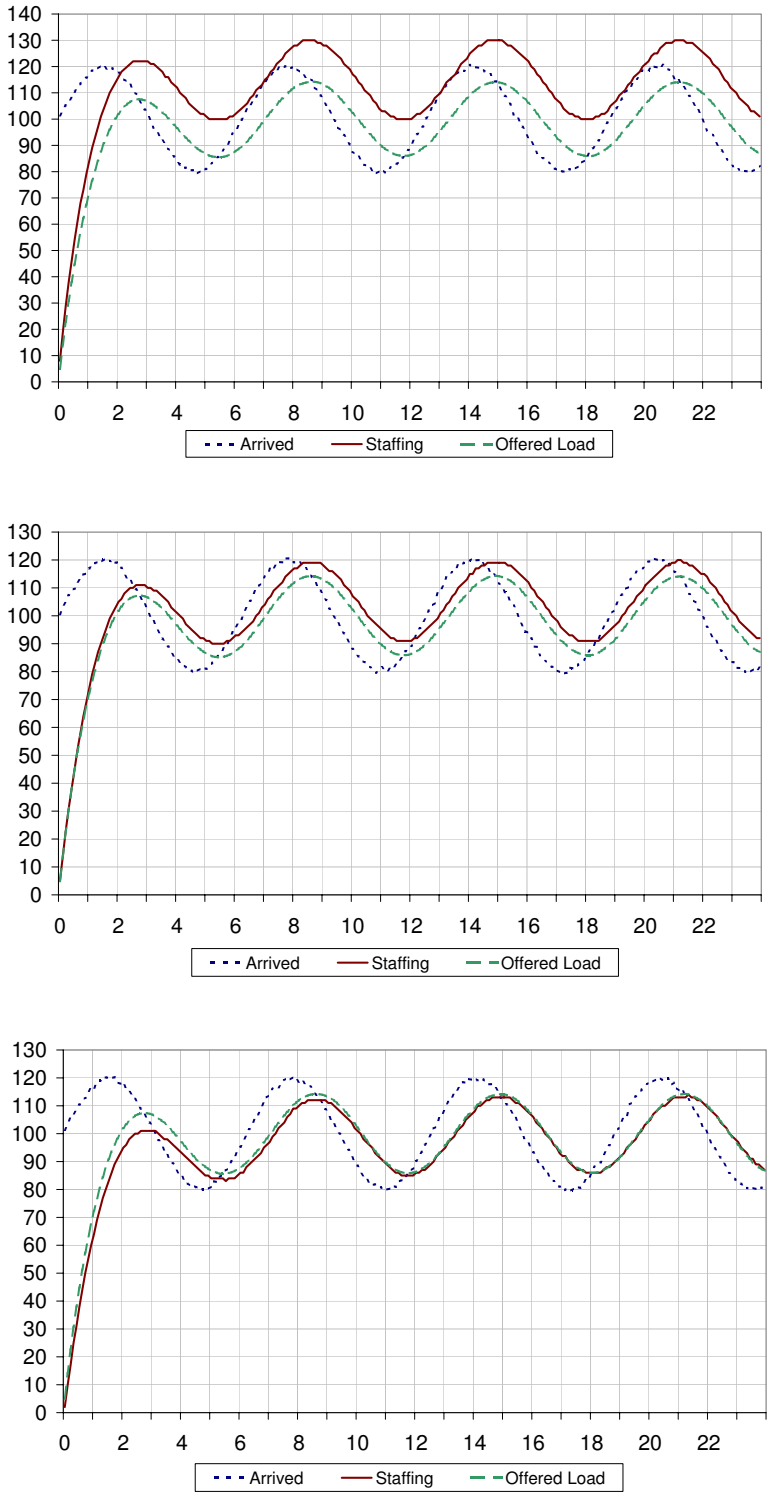


7. The Time-Varying Erlang-C Model

For comparison with the experiments for the time-varying Erlang-A ($M_t/M/s_t + M$) model in §5, we now show the performance of ISA for the same system described in §5 only without abandonment (with infinite patience) - the $M_t/M/s_t$ or time-varying Erlang-C model. As expected, the required staffing levels are higher than with abandonment, for all target delay probabilities; compare Figure 25 with Figure 3 in Section 5. For example, for $\alpha = 0.5$, the maximum staffing level becomes about 120 instead of 115.

For both the Erlang-A and Erlang-C models, the ISA staffing level decreases as the target delay-probability increases (as the performance requirement becomes less stringent) However, the staffing tends to coincide with the offered load in the Erlang-C example only in the ED regime, when $\alpha = 0.9$, as opposed to in the QED regime, when $\alpha = 0.5$. That illustrates how abandonment allows greater efficiency, while still meeting the delay-probability target.

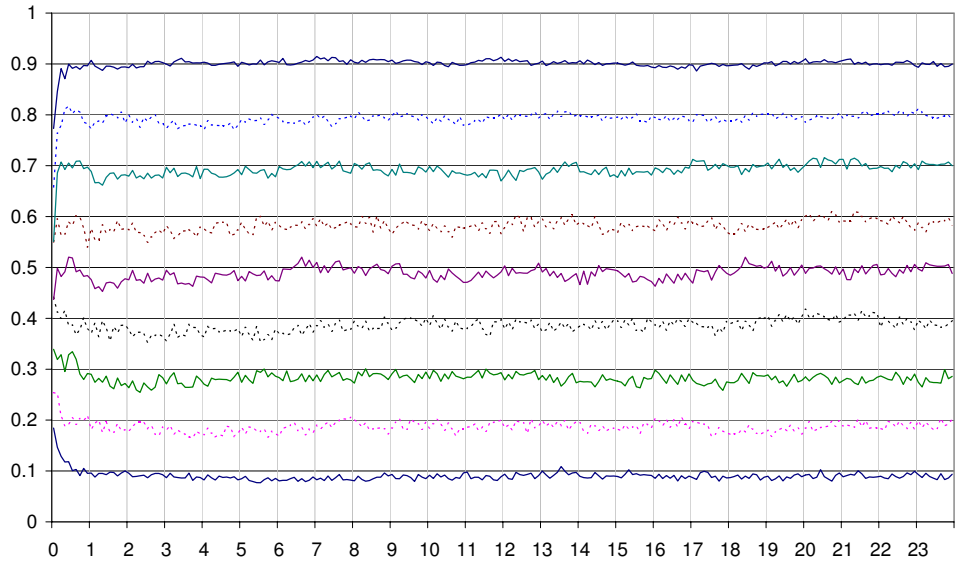
Figure 25: **The final staffing function found by ISA for the time-varying Erlang-C example with three different delay-probability targets: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.5$ (QED), (3) $\alpha = 0.9$ (ED)**



7.1. Time-Stable Performance

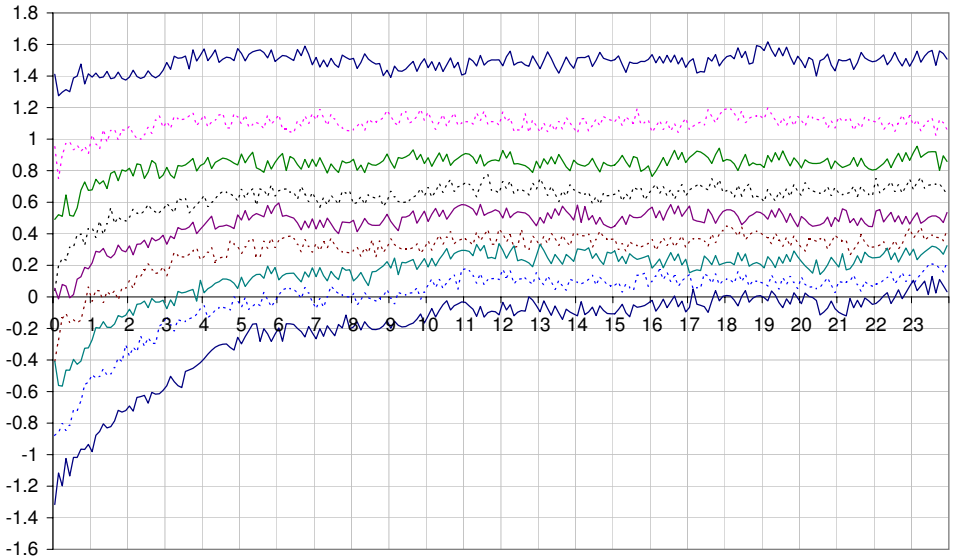
As before, we achieve accurate time-stable delay probabilities when we apply the ISA; see Figure 26, where again we consider target delay probabilities $0.1, 0.2, \ldots, 0.9$.

Figure 26: **Delay probability summary for the Erlang-C example**



The empirical service quality β_t^{ISA} is stabilizing as well, as can be seen from Figure 27, which shows results for the same 9 target delay probabilities. As in Figure 9, the empirical service

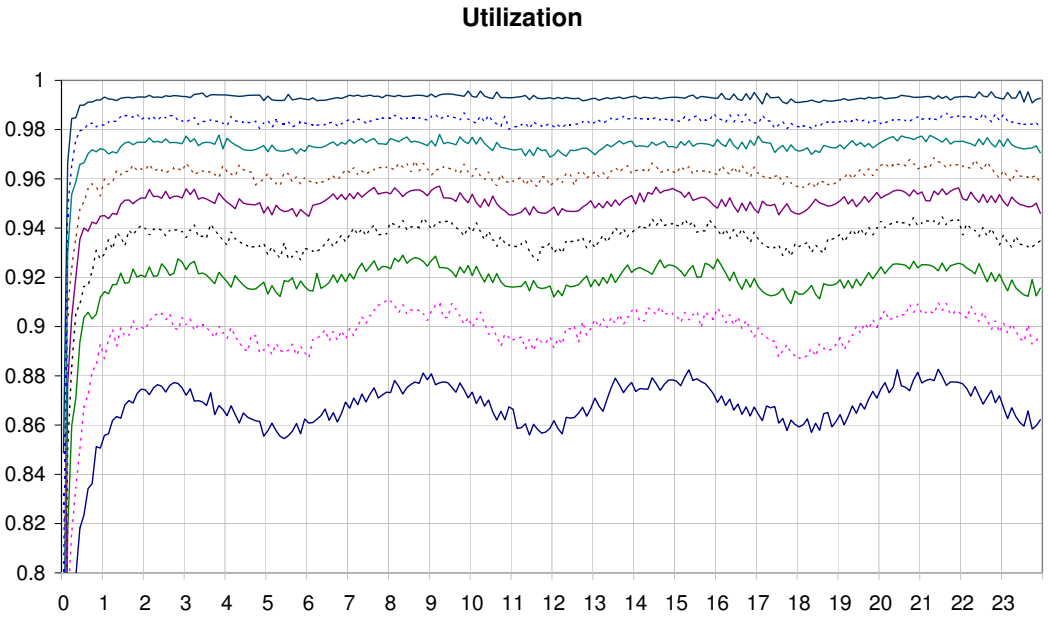
Figure 27: **Implied service quality β summary for the Erlang-C example** (The implied service quality decreases as α increases through the values 0.1, 0.2, ..., 0.9.)



quality decreases as the target delay probability increases. However, the empirical service quality β_t^{ISA} stabilizes at a much slower rate, especially for lower values of β (larger values of α). (The approach to steady-state is known to be slower in heavy traffic.) Nevertheless, the steady-state values can be seen at the right in Figure 27.

Without abandonment the system is more congested, but still congestion measures remain relatively stable. That is just as we would expect, since the time-dependent Erlang-C model is precisely the system analyzed in Jennings et al. (1996). Corresponding plots for other performance measures appear in Figures 28, 29, 30 and 31.

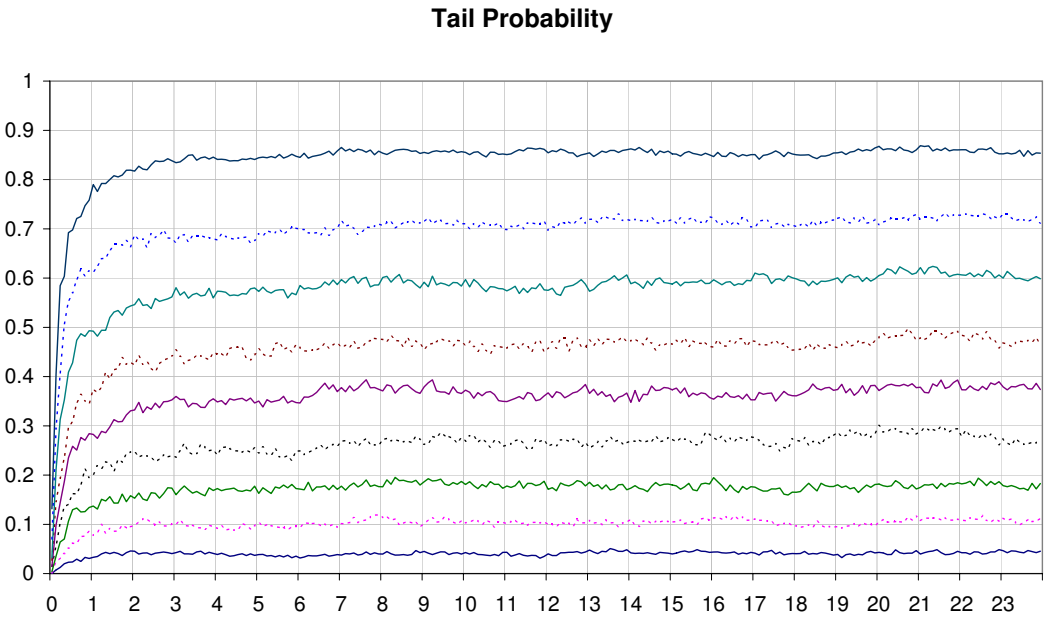
Figure 28: **Utilization summary for the Erlang-C example**



As stated in Section 5, precise explanations and definitions of the performance measures are given in Section 4.3.

Figures 27 and 30 show that here the time until system reaches (dynamic) steady-state is much longer compared to a system with abandonment. In fact, in Figure 30 steady-state was

Figure 29: **Tail probability summary for the Erlang-C example**



not yet reached after 24 time-units (the full day).

Figure 30: **Mean queue length and waiting time in the Erlang-C model with target α=0.5**

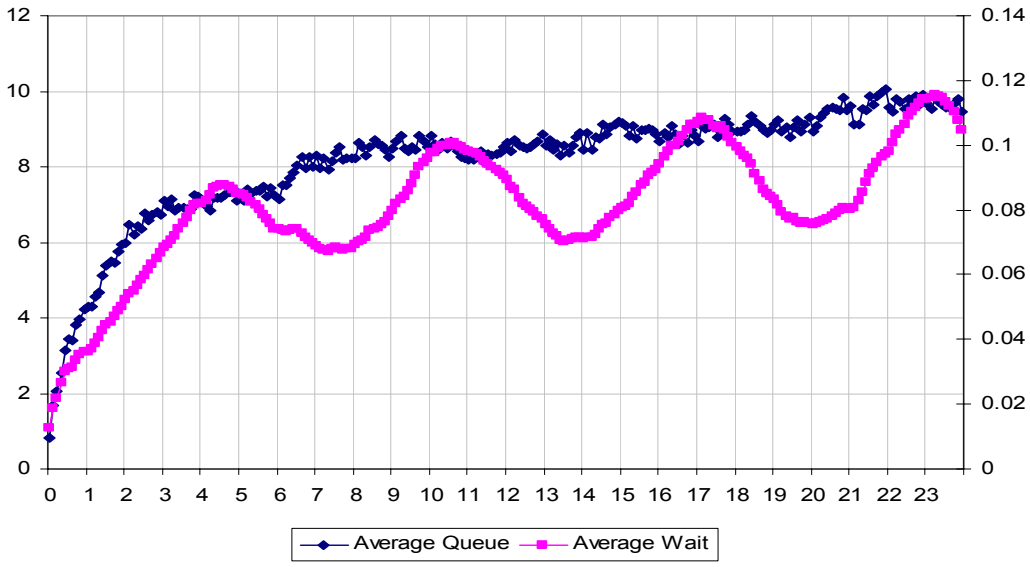
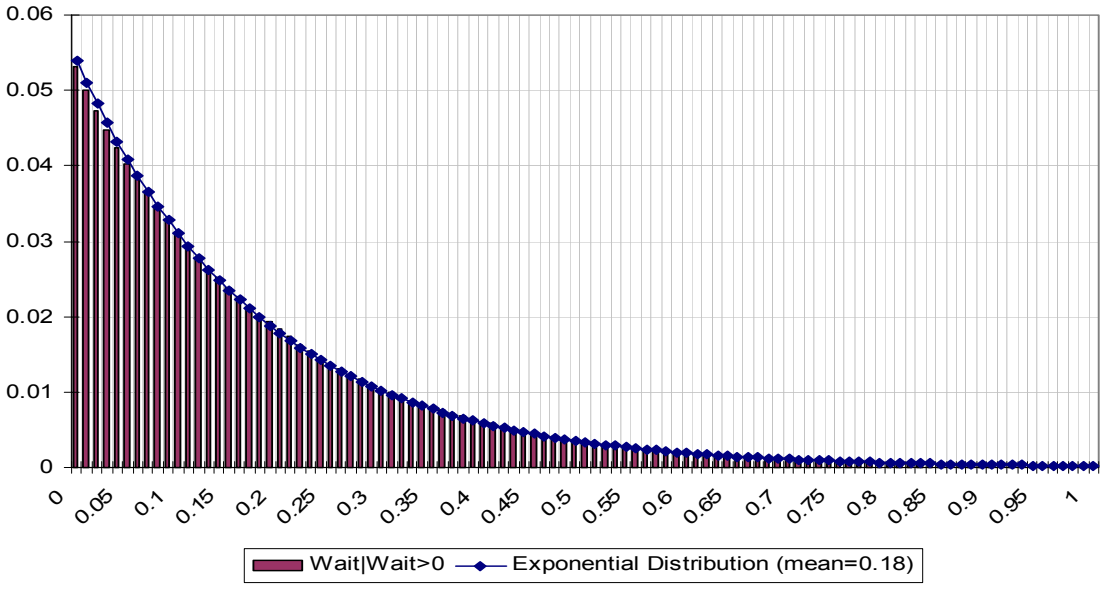


Figure 31: **The conditional distribution of the waiting time given delay in the Erlang-C model with target α=0.5**



7.2. Validating the Square-Root-Staffing Formula

Just as for the time-varying Erlang-A model, we want to validate the square-root-staffing formula in (2.7). We thus repeat the various experiments we did in §5. Recall that, for the *stationary* $M/M/s$ queue, the conditional waiting-time $(W \mid W > 0)$ is (exactly) exponentially distributed. The empirical conditional waiting-time distribution given wait, in our *time-varying* queue and over *all* customers, also fits the exponential distribution very well; see Figure 31. The mean of the plotted exponential distribution was taken to be the overall average waiting time of those who were actually delayed during $[0, T]$.

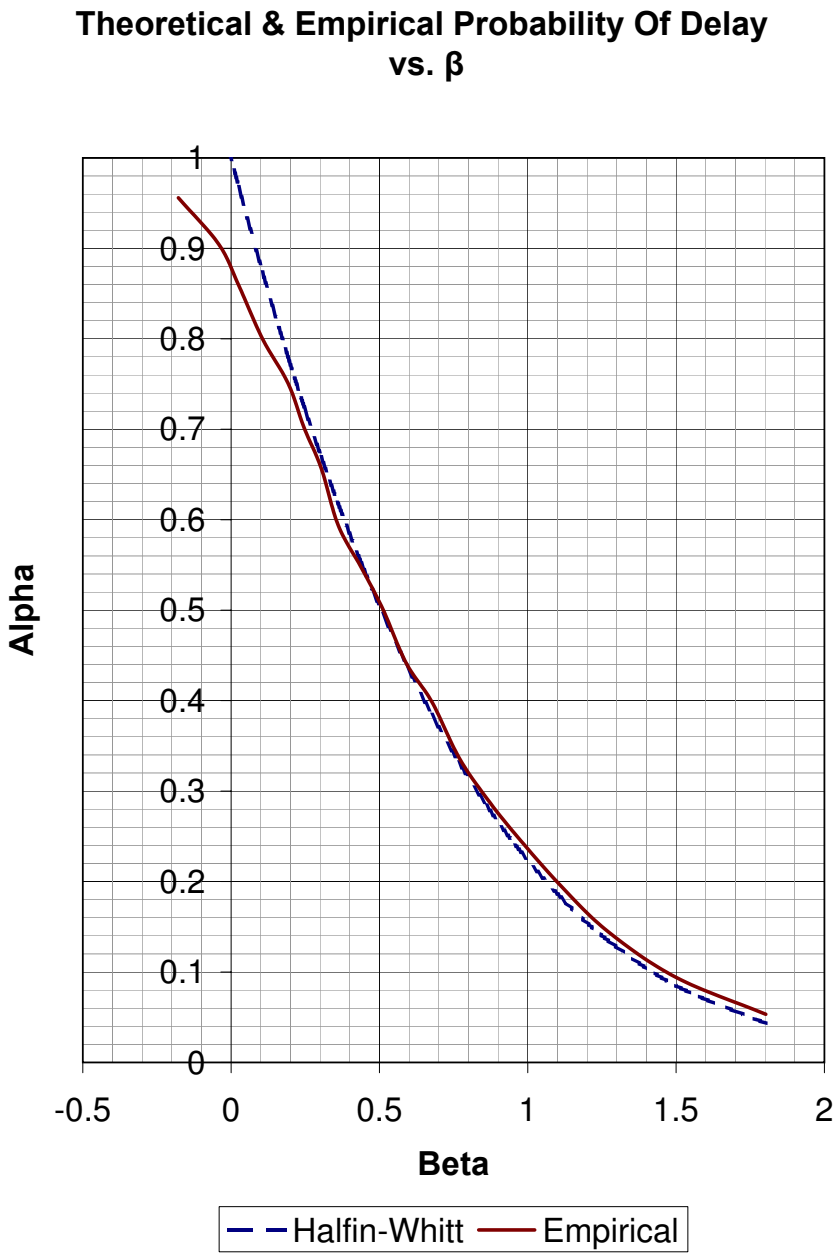
Here, the relation between α and β is compared with the **Halfin-Whitt function** from Halfin and Whitt (1981), namely,

$$P(delay) \equiv \alpha \equiv \alpha(\beta) \approx \left[1 + \beta \cdot \frac{\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \quad 0 < \beta < \infty \ , \tag{7.1}$$

where ϕ is again the pdf associated with the standard normal cdf Φ. The Halfin-Whitt function in (7.1) is obtained from the Garnett function in (5.4) by letting $\theta \to 0$.

Just as we use the Garnett function to relate the target delay probability α to the quality-of-service parameter β in the square-root-staffing formula in (2.7) for the $M_t/M/s_t + M$ model, so we use the Halfin-Whitt function to relate α to β in the square-root-staffing formula in (2.7) for the $M_t/M/s_t$ model. And that essentially corresponds to the refinement performed in Section 4 of Jennings et al. (1996). The results in Figure 32 are again remarkable.

Figure 32: **Comparison of empirical results with the Halfin-Whitt approximation for the Erlang-C example**

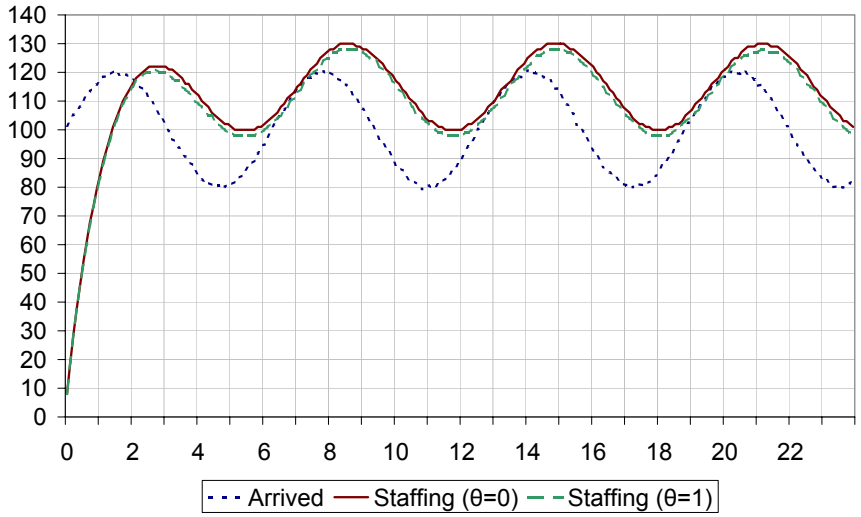


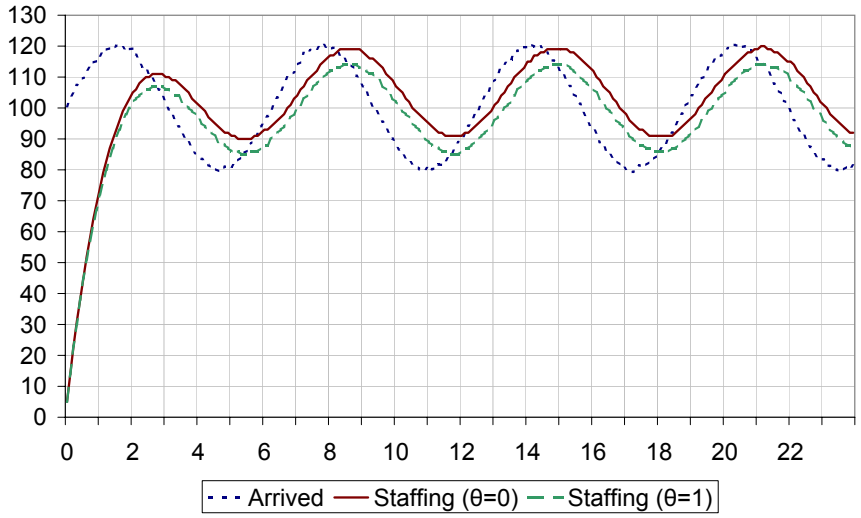
7.3. Benefits of Taking Account of Abandonment

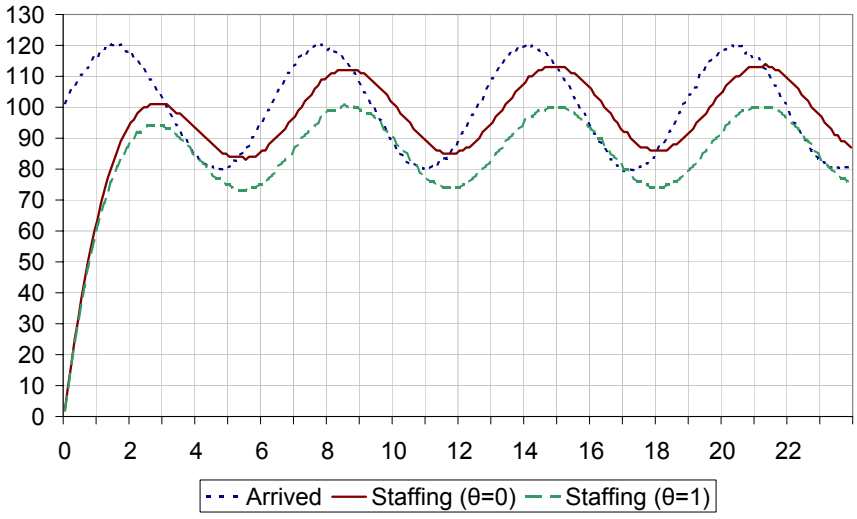
We now show the benefit of staffing a system taking account of abandonment, assuming that abandonment in fact occurs. When compared to the model without abandonment, abandonment in the model reduces the required staff. In Figure 33 we show the difference between staffing levels for the two models introduced in §5 and §7, in the three regimes of operation: QD, QED and ED.

It is natural to quantify the savings of labor by the area between the curves. In this case, the savings in labor, had one used $\theta = 1$, is 46.5 time units when $\alpha = 0.1$, 113.3 when $\alpha = 0.5$, and 256.4 when $\alpha = 0.9$. It may perhaps be better to quantify savings by looking at the savings of labor per shift. Dividing the saving in time-units by the number of time-units they are taken over, we come up with savings of about 2, 5 and 12 servers per shift, for $\alpha = 0.1, 0.5, 0.9$ respectively. The labor savings increases as α increases.

Figure 33: **Staffing levels with and without customer abandonment ($\theta = 0$ and $\theta = 0$): (1) $\alpha = 0.1$ (2) $\alpha = 0.5$ (3) $\alpha = 0.9$**







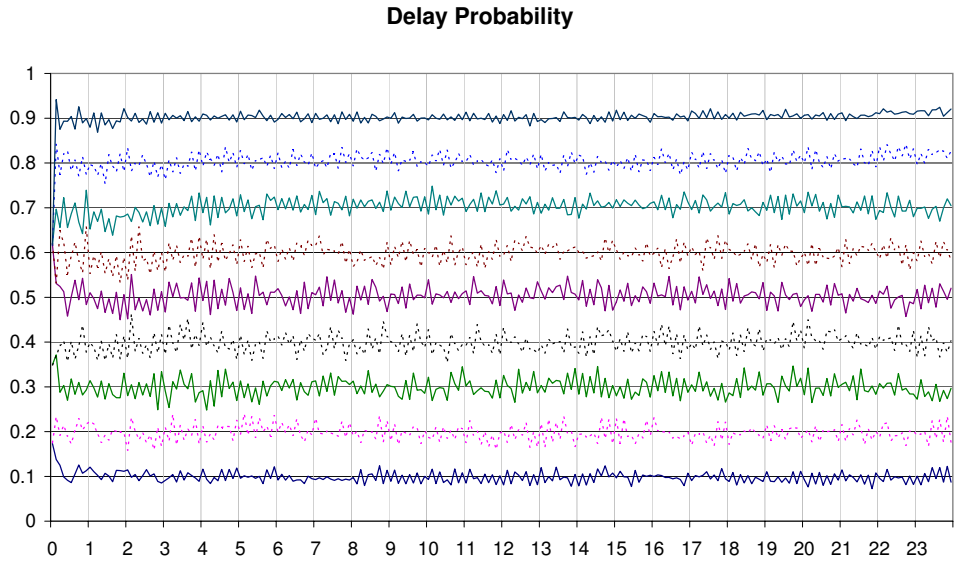
8. The Challenging Example from Jennings et al.

In this section, we consider the "challenging example" presented in Jennings et al. (1996). It is a time-varying Erlang-C model (no abandonment), with exponential service times having mean 1 and a nonhomogenous Poisson arrival process with the sinusoidal arrival-rate function $\lambda(t) = 30 + 20 \cdot \sin(5 \cdot t)$. We want to see how ISA performs on this same example.

This example is not greatly different from the Erlang-C example we have just considered in §7, but note that the frequency is 5 times greater here or, equivalently, the sinusoidal cycle is five times shorter. Thus the fluctuation in the arrival rate is even greater than in the example we have considered.

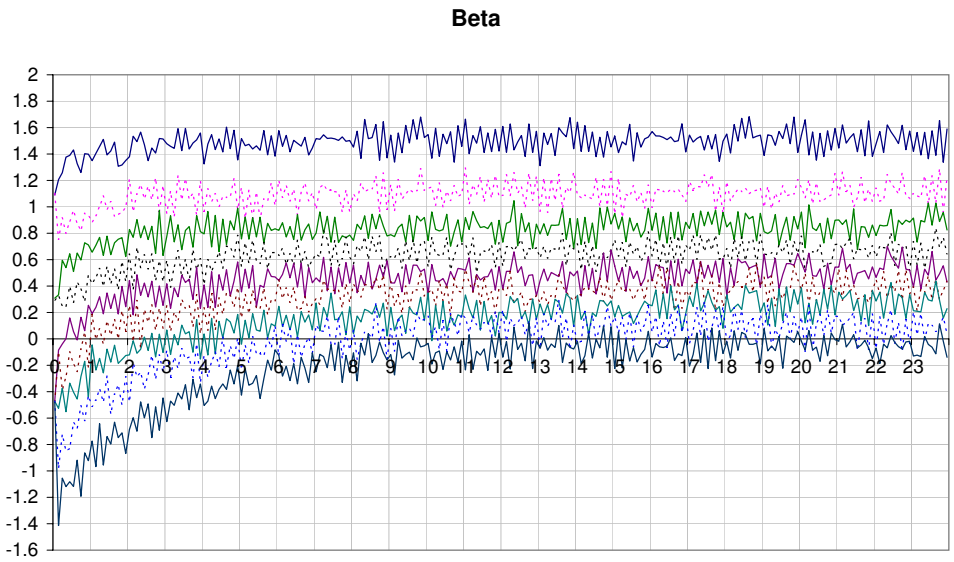
Figures 34 and 35 show that ISA also achieves time-stable performance for this example, for the full range of target delay probabilities, ranging from 0.1 to 0.9, just as before.

Figure 34: **Delay probability summary for the challenging example**



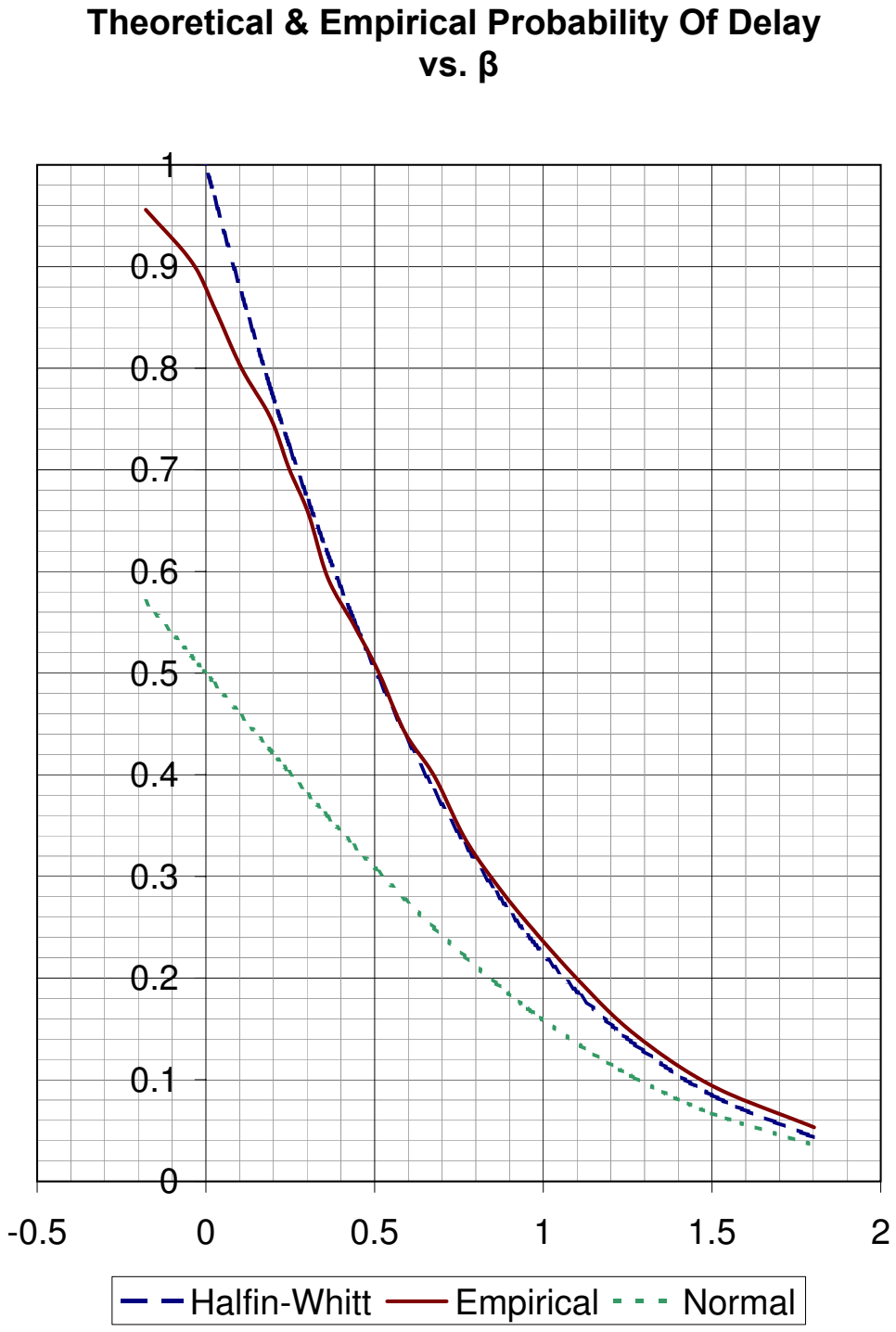
We now compare the empirical results with the Halfin-Whitt and normal approximations, paralleling Figures 11 and 32. We do so for this example below in Figure 36. Again the results are spectacular. In Figure 36 we use the Halfin-Whitt function in (7.1), just as in Figure

Figure 35: **Implied empirical quality of service β_t^{ISA} summary for the challenging example**



32. We also include the normal tail probability in (**??**), because that is the direct normal approximation used by Jennings et al. (1996), before they apply their refinement in their Section 4. That refinement is equivalent to working directly with the Halfin-Whitt function, as we noted before.

Figure 36: **Comparison of empirical results with the Halfin-Whitt and normal approximation for the challenging example**



9. The Time-Varying Erlang-A Model with More and Less Patient Customers

We now return to the time-varying Erlang-A model ($M_t/M/s_t + M$) considered in Section 5, except we change the patience parameter, i.e., the individual abandonment rate θ.

9.1. More and Less Patient Customers

We consider two new cases (both with $\mu = 1$: $\theta = 0.2$; then customers are **very patient**, since they are willing to wait, on average, five times the average service time; and $\theta = 5.0$; then customers are **very impatient**, since they are willing to wait, on average, only one-fifth of the average service time.

The performance of ISA is essentially the same as for the previous case with $\theta = 1.0$. We compare the staffing levels for these alternative environments, for the three regimes QD ($\alpha = 0.1$), QED ($\alpha = 0.5$), and ED ($\alpha = 0.9$) in Figure 37 below. In both these new cases, the target delay probabilities were achieved quite accurately for all target delay probabilities ranging from $\alpha = 0.1$ to $\alpha = 0.9$; see Figure 38. The implied empirical quality of service β_t^{ISA} defined in (5.3) is also stable, just as with $\theta = 1.0$; see Figure **??**. We compare the time-dependent abandonment $P_t(Ab)$ in these two scenarios in Figure 40. Note that the gap between the required staffing levels in the two cases - $\theta = 0.2$ and $\theta = 5.0$ - grows as the delay-probability target α increases, being quite small when $\alpha = 0.1$, but being very dramatic when $\alpha = 0.9$.

We compare the empirical (α, β) pairs produced by ISA to the Garnett function in (5.4) for these two cases in Figure 41. We are no longer surprised to see that the fit is excellent.

From all our studies of ISA, we conclude that for the time-varying Erlang-A model we can always use the MOL approximation, here manifested in the square-root-staffing formula in (2.7), obtaining the required service quality β from the target delay probability α by using the inverse of the Garnett function in (5.4), which reduces to the Half-Whitt function in (7.1) when $\theta = 0$. To see how the Garnett functions look, we plot the Garnett function for several values of the ratio $\mathbf{r} \equiv \theta/\mu$ in Figure 42 below.

Figure 37: **Staffing for time-varying Erlang-A with more patient ($\theta = 0.2$) and less patient ($\theta = 5.0$) customers: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.9$ (ED), (3) $\alpha = 0.5$ (QED)**

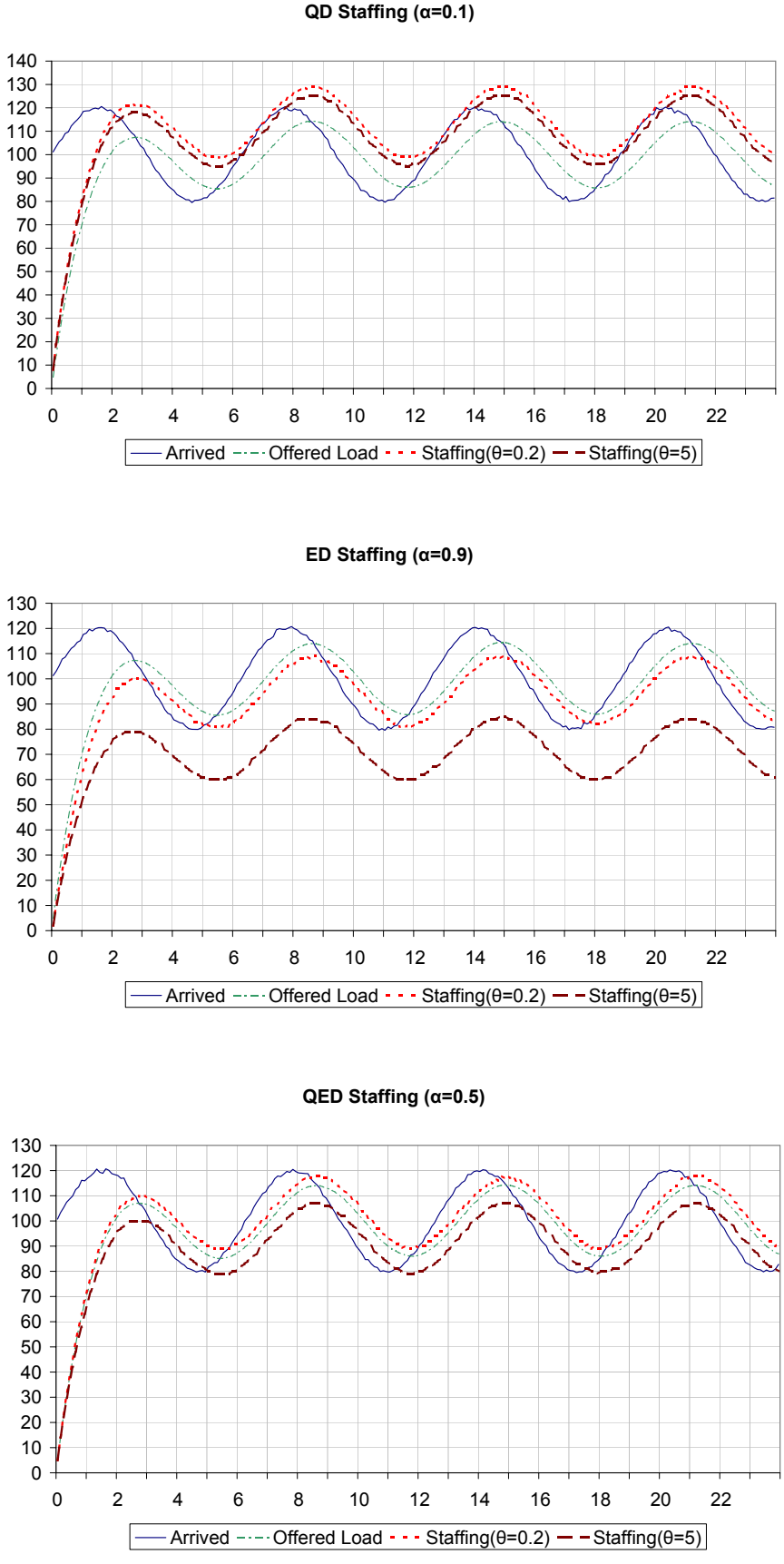
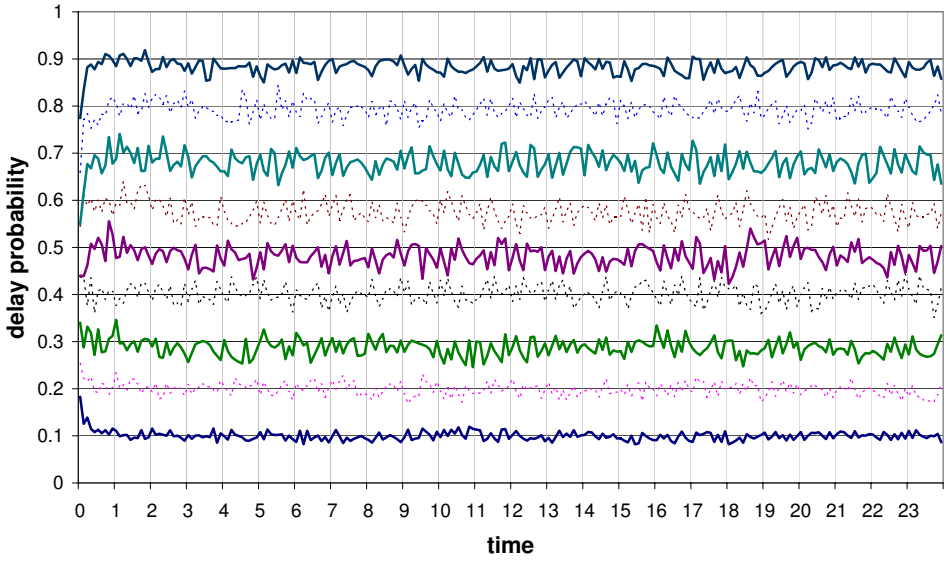


Figure 38: **Delay probabilities for the time-varying Erlang-A example with the new patience parameters: (1) θ=5 (2) θ=0.2**



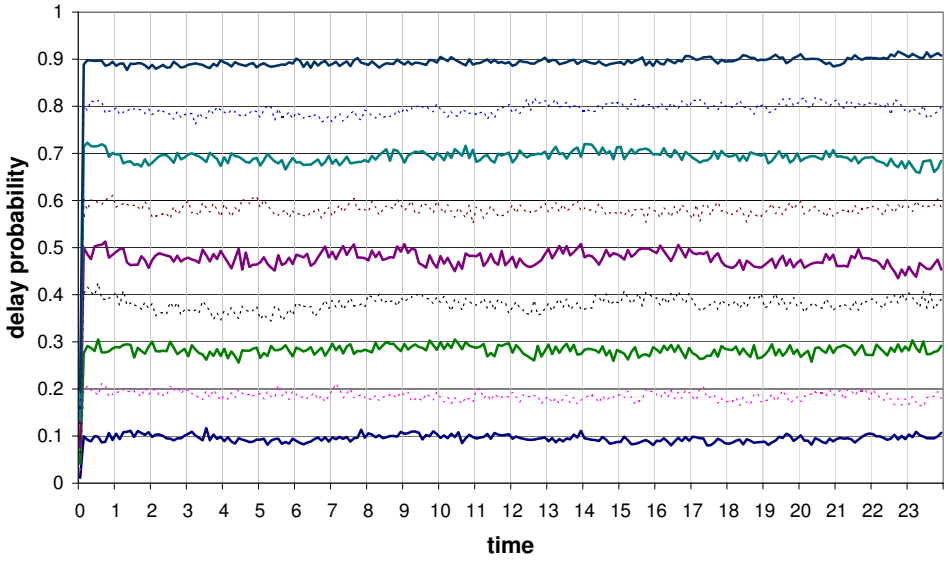
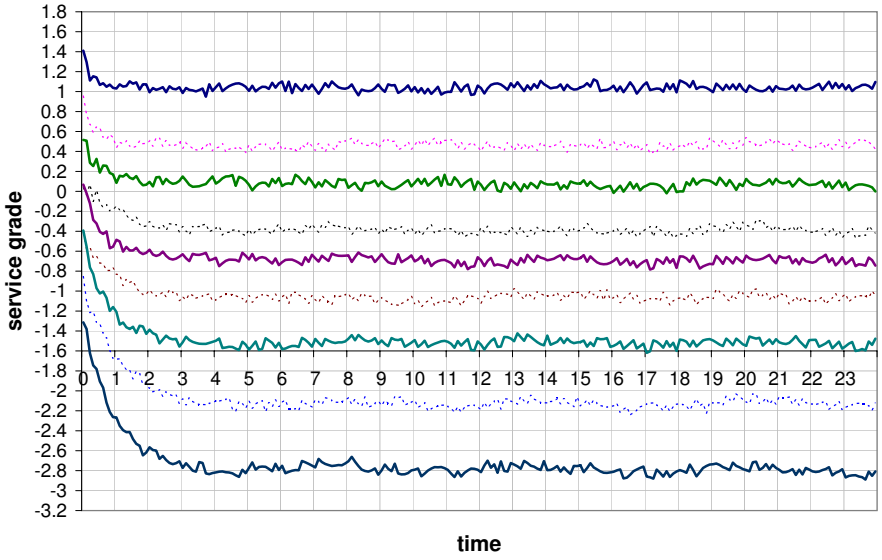


Figure 39: **Implied empirical quality of service β_t^{ISA} for the time-varying Erlang-A example with the new patience parameters: (1) θ=5 (2) θ=0.2**



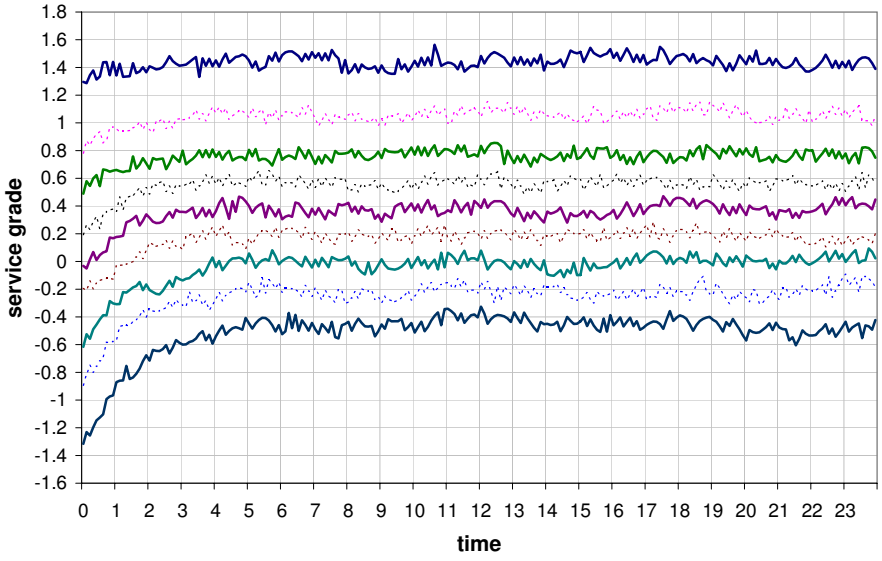
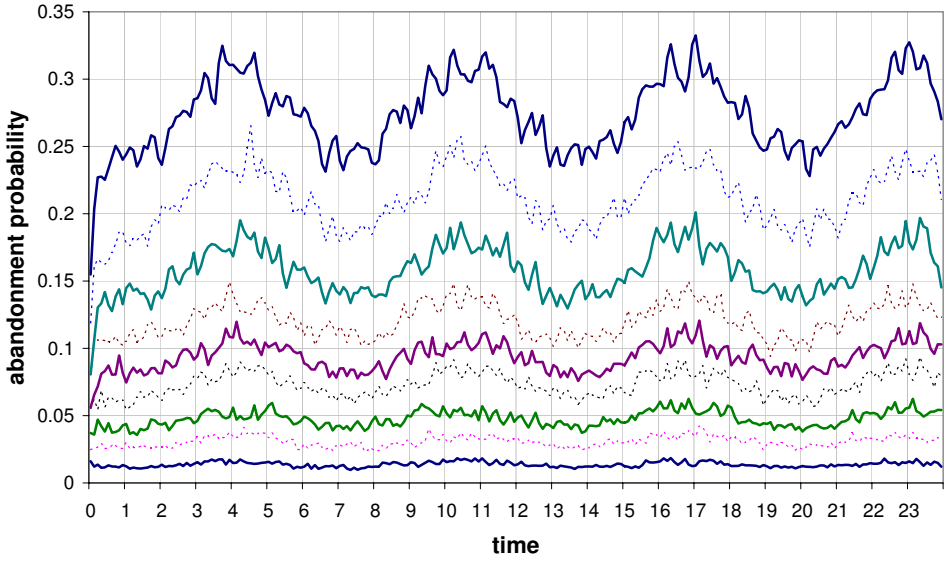


Figure 40: **Abandonment probabilities for the time-varying Erlang-A example with the new patience parameters: (1) θ=5 (2) θ=0.2**



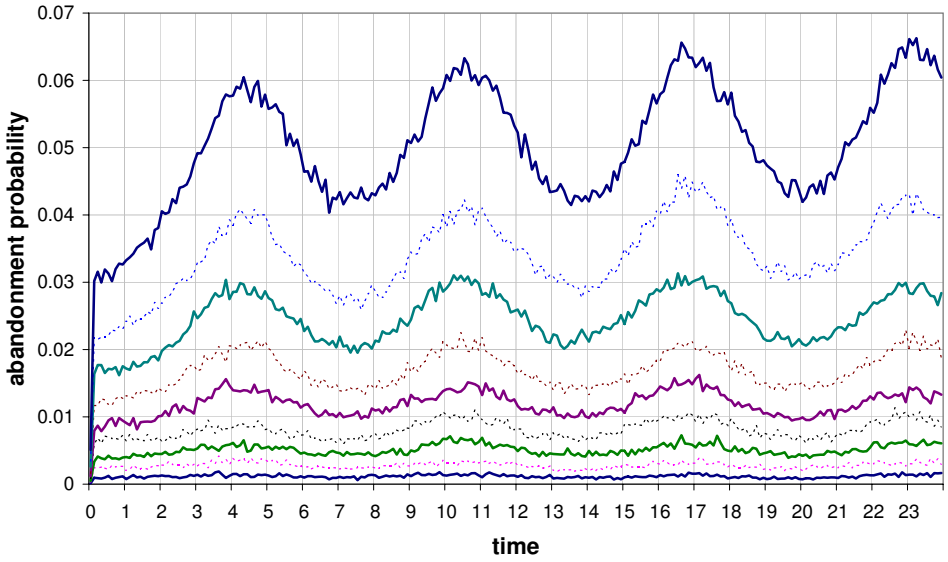


Figure 41: **Comparison of the empirical results from ISA with the Garnett approximation for the time-varying Erlang-A example with the new patience parameters: θ=5 and θ=0.2**

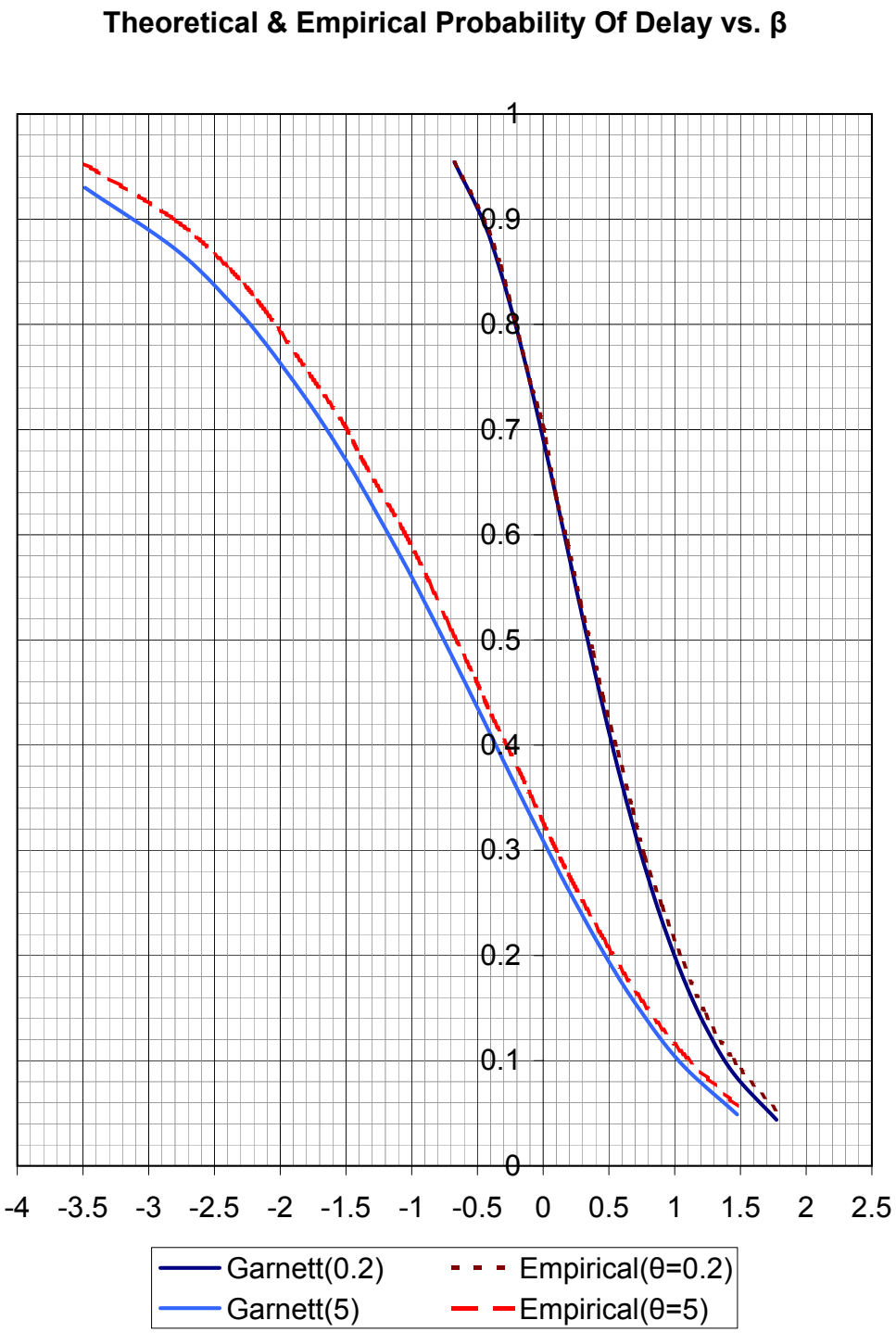
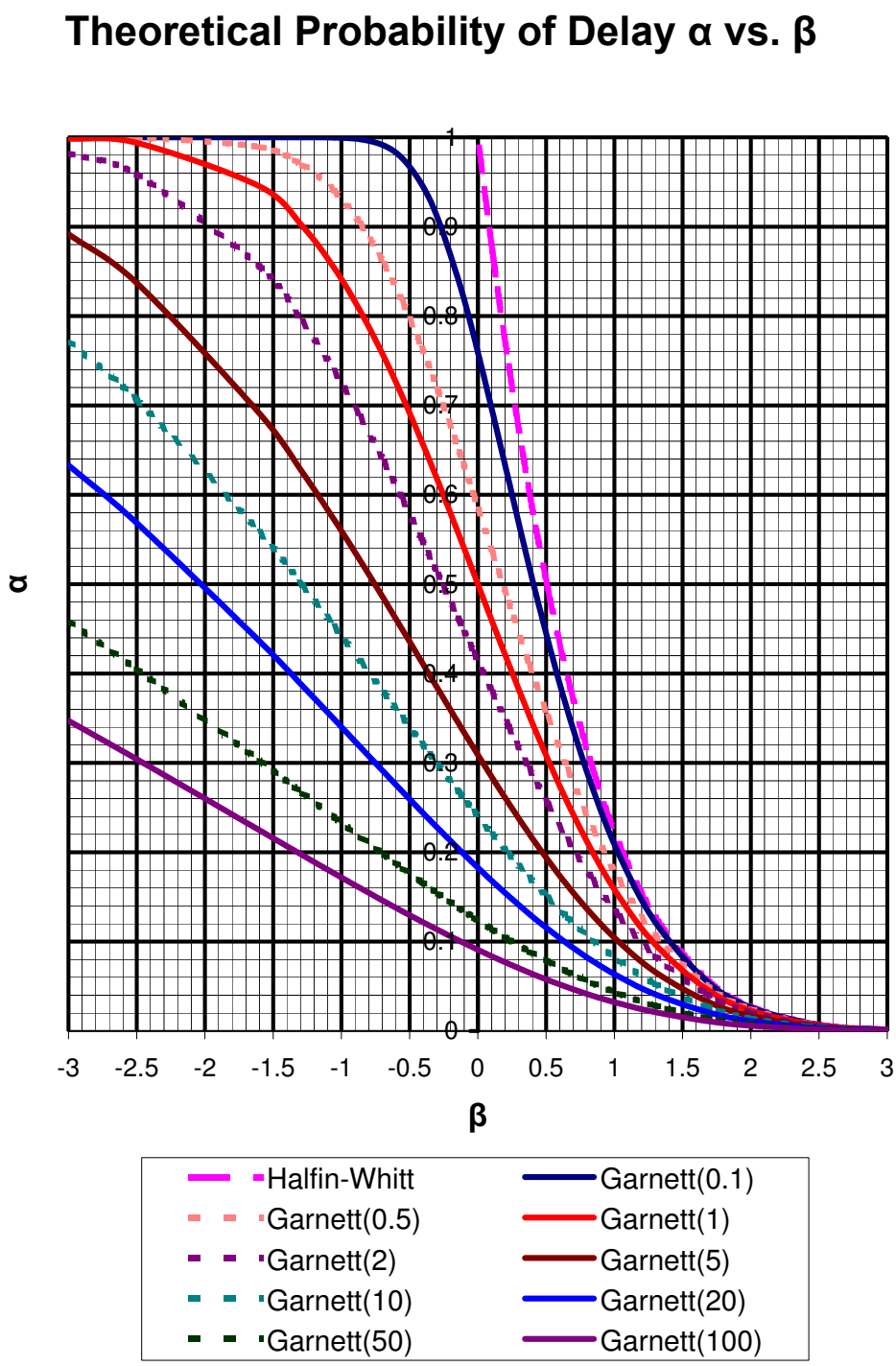


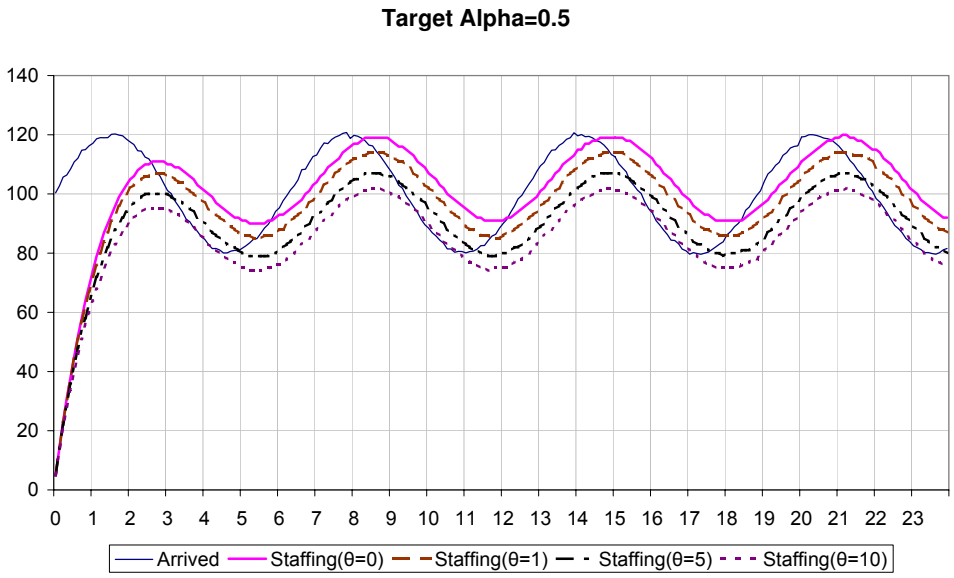
Figure 42: **The Halfin-Whitt/ Garnett functions**



9.2. Benefits of Taking Account of Abandonment Again

Following §7.3, we now expand our comparison of staffing levels for (im)patience distribution with parameters $\theta = 0, 1, 5, 10$. Clearly, the required staffing level decreases as θ increases, bringing additional savings. In Figure 43 we show the comparison for delay probability $\alpha = 0.5$, which we consider to be a reasonable operational target.

Figure 43: **Staffing levels for the time-varying Erlang-A example for a range of (im)patience parameters**



Here, the labor savings is: 113.3 time units for $\theta = 1$, 270 time units for $\theta = 5$, and 386 time units for $\theta = 10$. The corresponding savings in workers per shift are about 5, 12 and 18 servers, for $\theta = 1, 5, 10$, respectively.

9.3. Non-Exponential Service Times

In addition to the time-varying Erlang-C and Erlang-A examples, we also ran experiments with different service-time distributions, such as deterministic and log-normal. The ISA was successful in achieving the desired target delay probability, and results showed time-stable performance, compatible with stationary theory, similar to here. For the case of deterministic service times, theory was taken from Jelenkovic, Mandelbaum and Momcilovic (2004).

10. Theoretical Support in the Case $\theta = \mu$

In one special case, we can analyze the time-dependent Erlang-A model (i.e., the $M_t/M/s_t+M$ model) in considerable detail. That is the case we considered in Section 5, in which the individual service rate μ equals the individual abandonment rate θ. In this section, let θ and μ be fixed with $\theta = \mu$, but here we do not set these equal to 1.

10.1. Connections to Other Models

In one special case, we can analyze the $M_t/M/s_t + M$ model in considerable detail. That is the case we considered in §5 and §6, in which $\theta = \mu$. (As in §5, we let those both be 1.) With the condition $\theta = \mu$, it is easy to relate the $M_t/M/s_t + M$ model to, first, the corresponding $M_t/M/\infty$ model with the same arrival-rate function and service rate and, second, a corresponding family of steady-state distributions of stationary $M/M/s + M$ models, indexed by t, with the same service and abandonment rates, but with special arrival rate that depends on time t.

Let $\{s_t : t \geq 0\}$ be an arbitrary staffing function. For simplicity, assume that all systems start empty in the distant past (at time $-\infty$). By having $\lambda(t) = 0$ for $t \leq t_0$, we can start arrivals at any time t_0. The first observation is that, for any arrival-rate function $\{\lambda(t) : t \geq 0\}$ and any staffing function $\{s_t : t \geq 0\}$, the stochastic process $\{N_t : t \geq 0\}$ in the $M_t/M/s_t + M$ model with $\theta = \mu$ has the same distribution (finite-dimensional distributions) as the corresponding process $\{N_t^\infty : t \geq 0\}$ in the $M_t/M/\infty$ model, because the birth and death rates are the same.

The second observation is that, for both these models, the individual random variables N_t and N_t^∞ have the same Poisson distribution as the steady-state number in system $N_\infty^{(t)}$ in the corresponding stationary model with arrival rate m_t^∞.

10.2. Waiting times and abandonment probabilities.

Let W_t be the *virtual waiting time* at time t (until service, i.e., the waiting time in queue that would be spent by an infinitely patient customer arriving at time t), and let P_t^{ab} be the *virtual abandonment probability* at time t (i.e., the probability of abandonment for an arrival that would occur at time t), both in the $M_t/M/s_t + M$ model. These quantities are considerably more complicated than N_t.

Even though it is difficult to evaluate the full distribution of W_t, we can immediately evaluate the virtual delay probability, because it clearly depends only on what the customer

encounters upon arrival at time t. Hence, we have

$$P(W_t > 0) = P(N_t \geq s_t) = P(N_t^\infty \geq s_t) = P(Poisson(m_t^\infty) \geq s_t) \ , \tag{10.1}$$

where m_t^∞ is the offered load in (2.2), just as in (2.6), only here the infinite-server approximation is exact.

Next we observe that $P_t^{ab} = E[F(W_t)]$, where F is the time-to-abandon cdf, so that it suffices to determine the waiting-time distribution. Here is an important initial observation: Conditional on the event that $W_t > 0$, whose probability we have characterized above, W_t is distributed (exactly) as the first passage time of the (Markovian) stochastic process $\{N_u : u \geq t\}$ from the initial value N_t encountered at time t down to the staffing function $\{s_u : u \geq t\}$, provided that we ignore all future arrivals after time t. In other words, W_t is distributed as the first passage time of the pure-death stochastic process with state-dependent death rate N_u for $u \geq t$ down from the initial value N_t to the curve $\{s_u : u \geq t\}$. As a consequence, the distribution of W_t and the value of P_t^{ab} depend on only N_t and the future staffing levels, i.e., $\{s_u : u \geq t\}$. The time-dependent arrival-rate function contributes nothing further.

It is easy to see that we can establish stochastic bounds on the distribution of W_t if the staffing level is monotone after time t: then setting $s_u = s_t$ for all $u \geq t$ will yield a bound. We can go further based on this observation if we make approximations. If the number of servers is large, then W_t will tend to be small, so that it is often reasonable to make the approximation $s_u \approx s_t$ for all $u > t$. We make this approximation, not because the staffing level should be nearly constant for all u after t, but because we think we only need to consider times u slightly greater than t.

If the future-staffing-level approximation held as an equality, then we would obtain the following approximations as equalities: $W_t \approx W_\infty$ and $P_t^{ab} \approx P_\infty^{ab}$, where the constant staffing level in the stationary $M/M/s + M$ model on the righthand sides is chosen to be s_t and the constant arrival rate is chosen to be λ_t^{MOL} in (2.8). Given these approximations, we can use established results for the stationary $M/M/s + M$ model, e.g., as in Garnett et al. (2002) and Whitt (2005). Algorithms to compute the (exact) distribution of W_∞ are given there, including the corresponding conditional distributions obtained when we condition on whether or not the customer eventually is served.

10.3. Asymptotic Time-Stability in the Many-Server Heavy-Traffic Limit

In this subsection we turn to an issue not included in the main paper. As in the literature for stationary models, e.g., Garnett et al. (2002), important insight can be gained by considering

many-server heavy-traffic limits. That is achieved for our $M_t/M/s_t + M$ model, by considering a sequence of models indexed by n, where the arrival-rate function is allowed to depend upon n. We can leave the service rate and abandonment rate unchanged, independent of n (and t). Thus, for each n, we have arrival-rate function $\lambda_n \equiv \{\lambda_n(t) : t \geq 0\}$. As in the stationary context, we want to let the arrival rate increase as $n \to \infty$. However, now we need to carefully specify how the entire function λ_n increases. Since we are staffing in response to the arrival rate, we do not need to make any direct assumptions about the staffing levels s_t. We will assume that we staff according to the square-root-staffing formula (2.7) with a fixed target delay probability α. We then want to determine when that yields asymptotically time-stable performance.

As an initial condition, we want to assume that $\lambda_n(t) \to \infty$ as $n \to \infty$ for each t, but we will need more than that. From the analysis so far, it is clear that we need $m_{t,n} \to \infty$, where $m_{t,n}$ is the time-dependent mean number in the n^{th} $M_t/M/\infty$ model. However, that actually is not enough to get asymptotic time-stability for quantities such as the mean virtual waiting time $E[W_t]$ and the virtual abandonment probability P_t^{ab}.

To proceed, we exploit the approximations above. From the approximation for the mean, we obtain the associated approximation

$$E[W_t] \approx E[W_\infty] \tag{10.2}$$

where the constant staffing level in the stationary $M/M/s + M$ model on the righthand side is chosen to be s_t and the constant arrival rate is chosen to be $\hat{\lambda}(t) \equiv \mu m_t$ in (2.8).

Now we observe that previous heavy-traffic limits for the Erlang-A model in the QED regime, Theorems 3 and 4 of Garnett et al. (2002), imply that

$$\sqrt{m_t} P_t^{ab} \to \eta \quad \text{and} \quad \sqrt{m_t} E[W_t] \to \frac{\eta}{\theta} \tag{10.3}$$

as $m_t \to \infty$, where

$$\eta \equiv \alpha E[N(0,1) - \beta | N(0,1) > \beta] = \alpha \left(\frac{\phi(\beta)}{\Phi^c(\beta)} - 1 \right) > 0 \tag{10.4}$$

and $\theta = \mu$.

The important practical conclusion we deduce from (10.3) is that we see that $\sqrt{m_t} P_t^{ab}$ and $\sqrt{m_t} E[W_t]$ will be asymptotically constant (time-stable and nondegenerate) as m_t increases if we are in the QED regime. However, in general, consistent with Figure 7, the performance measures P_t^{ab} and $E[W_t]$ themselves need not be asymptotically time-stable. In order for

them to be asymptotically time-stable too, we need to impose extra conditions upon the mean function m_t itself.

We actual see the greatest departures from time-stability of P_t^{ab} and $E[W_t]$ for the $M_t/M/s_t+M$ model (e.g., in Figure 7) when the target delay probability is large. In those cases, it is evident that the system actually should be regarded as being in the ED regime, not the QED regime. From Garnett et al. (2002) and Whitt (2004), we can see the appropriate ED asymptotics, which also suggests that time-stability will not hold for the performance measures P_t^{ab} and $E[W_t]$, staffing as we have done. Moreover, it suggests that we might consider a different staffing method designed to achieve time-stable abandonment in the ED regime. In particular, ISA extends directly by changing the target performance measure from the delay probability to the abandonment probability. The performance of such alternative iterative-staffing algorithms is a topic for future research.

11. Algorithm Dynamics

In this section we establish the convergence of ISA for the $M_t/M/s_t + M$ model. In doing so, we disregard statistical error caused by having to estimate the delay probabilities associated with each staffing function in the simulation.

11.1. Sample-Path Stochastic Order

To prove convergence, we use sample-path stochastic order, as in Whitt (1981). We say that one stochastic process $\{N_t^{(1)} : 0 \le t \le T\}$ is stochastically less than or equal to another, $\{N_t^{(2)} : 0 \le t \le T\}$, in *sample-path stochastic order* and write

$$\{N_t^{(1)} : 0 \le t \le T\} \le_{st} \{N_t^{(2)} : 0 \le t \le T\} \ , \tag{11.1}$$

if

$$E\left[f\left(\{N_t^{(1)} : 0 \le t \le T\}\right)\right] \le_{st} E\left[f\left(\{N_t^{(2)} : 0 \le t \le T\}\right)\right] \tag{11.2}$$

for all nondecreasing real-valued functions f on the space of sample paths. We have ordinary stochastic order for the individual random variables $N_t^{(1)}$ and $N_t^{(2)}$ and write $N_t^{(1)} \le_{st} N_t^{(2)}$ if $E[f(N_t^{(1)})] \le E[f(N_t^{(2)})]$ for all nondecreasing real-valued functions on the real line; see Chapter 9 of Ross (1996) and Müller and Stoyan (2002). Clearly, sample-path stochastic order as in (11.1) implies ordinary stochastic order for the individual random variables for all t. For the convergence, we only need ordinary stochastic-order for each time t, but in order to get that, we need to properly address what happens before time t as well.

Here is the key stochastic-comparison property for the $M_t/M/s_t + M$ model:

Theorem 11.1. (stochastic comparison) *Consider the $M_t/M/s_t + M$ model on the time interval $[0,T]$, starting empty at time 0. If $r \geq 1$ and $s_t^{(1)} \leq s_t^{(2)}$ for all t, $0 \leq t \leq T$, or if $r \leq 1$ and $s_t^{(1)} \geq s_t^{(2)}$ for all t, $0 \leq t \leq T$, then*

$$\{N_t^{(1)} : 0 \leq t \leq T\} \leq_{st} \{N_t^{(2)} : 0 \leq t \leq T\} \ . \tag{11.3}$$

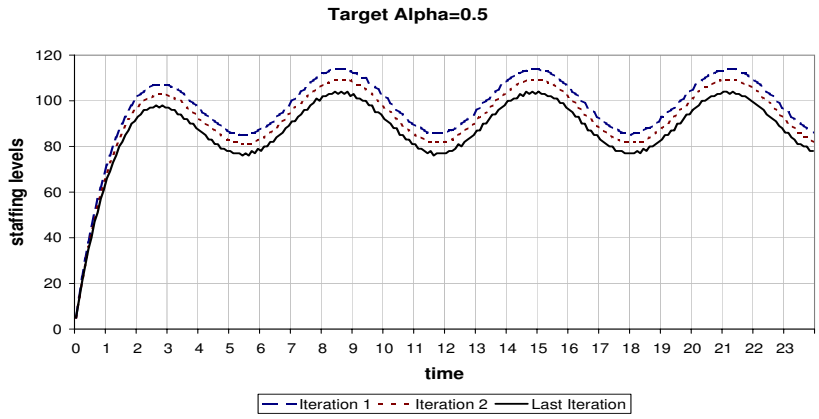
Proof. Here is the key fact: The death rates depend systematically on the number of servers s_t. When $r > 1$ ($r < 1$), the death rates at time t decrease (increase) as s_t increases. The ordering of the death rates in the two birth-and-death processes makes it possible to achieve the sample-path ordering. Indeed, we justify the relation (11.3) by constructing special versions of the two stochastic processes on the same underlying probability space so that the sample paths are ordered with probability 1. As discussed in Whitt (1981), and proved by Kamae et al. (1978), that special construction is actually equivalent to the sample-path stochastic ordering in (11.3). The sample-path ordering obtained ensures that a departure occurs in the lower process whenever it occurs in the upper process and the two sample paths are equal. To start the construction, we let the two processes be given identical arrival streams. Then we construct all departures (service completions or abandonments) from those of the lower process at epochs when the two sample paths are equal. Suppose that at time t the sample paths are equal: $N_t^{(1)} = N_t^{(2)} = k$. Then, at that t, the death rates in the two birth and death processes are necessarily ordered by $\delta_1(k) \geq \delta_2(k)$. We only let departures occur in process 2 when they occur in process 1, so the two sample paths can never cross over. When a departure occurs in process 1 with both sample paths in state k, we let a departure also occur in process 2 with probability $\delta_2(k)/\delta_1(k)$, with no departure occurring in process 2 otherwise. This keeps the sample paths ordered w.p. 1 for all t. At the same time, the two stochastic processes individually have the correct finite-dimensional distributions. ∎

11.2. Monotone and Oscillating Dynamics

The simulation experiments show that the way the staffing functions converge to the limit depends on the ratio $r \equiv \theta/\mu$: Whenever $r > 1$, we encounter monotone dynamics. Whenever $r < 1$, we encounter oscillating dynamics; and whenever $r = 1$, we encounter instantaneous convergence. As shown in §10, when $r = 1$, the number in system is independent of the staffing function, so we obtain convergence in one step.

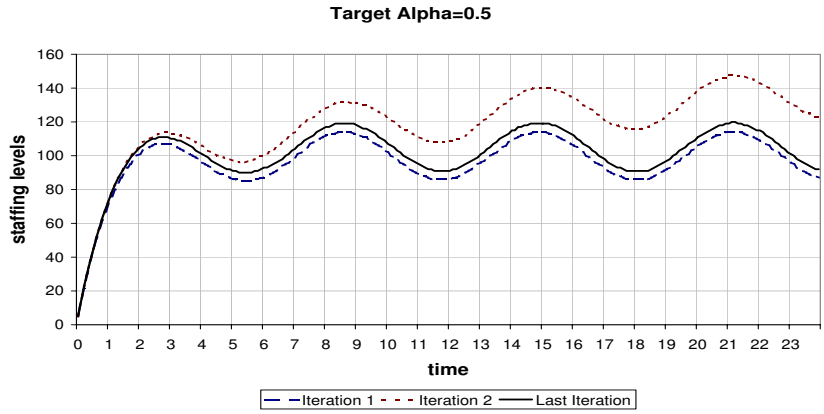
An example of the monotone dynamics is shown in Figure 44, where staffing levels are shown for the first two and final iterations for the model in §5 with $\mu = 1$ and $r = \theta = 10$. An example of the oscillating dynamics is shown in Figure 45, where staffing levels are shown

Figure 44: **Monotone algorithm dynamics for the model in §5 when $r = \theta = 10$: staffing levels in the 1^{st}, 2^{nd} and final iterations.**



for the first two and final iterations for the model in §5 with $\mu = 1$ and $r = \theta = 0$ (no abandonment).

Figure 45: **Oscillating algorithm dynamics for the model in §5 when $r = \theta = 0$: staffing levels in the 1^{st}, 2^{nd} and final iterations.**



11.3. Proof of Convergence

Theorem 11.2. (convergence) *Consider the $M_t/M/s_t + M$ model on the time interval $[0, T]$, starting empty at time 0. Suppose that we consider piecewise-constant staffing functions that only can change at multiples of $\Delta > 0$. Suppose that in each iteration n we can obtain the actual stochastic process $\{N_t^{(n)} : 0 \leq t \leq T\}$ associated with the staffing function $\{s_t^{(n)} : 0 \leq t \leq T\}$ (without statistical error). Suppose that $s_t^{(0)} = \infty$ for all t, $0 \leq t \leq T$.*

(a) If $r > 1$, then $s_t^{(n)} \leq s_t^{(m)}$ for all $n > m \geq 0$ and there exists a positive integer n_0 such that

$$s_t^{ISA} = s_t^{(n_0)} = s_t^{(n)} \quad \text{for all} \quad t \quad \text{and} \quad n \geq n_0 \ . \tag{11.4}$$

(b) If $r < 1$, then there exist 2 subsequences $\{s_t^{(2n)}\}$ and $\{s_t^{(2n+1)}\}$, such that $s_t^{(2n)} \downarrow s_t^{(even)}$ and $s_t^{(2n+1)} \uparrow s_t^{(odd)}$.

$$s_t^{(0)} \geq s_t^{(2n)} \geq s_t^{(2n+2)} \geq s_t^{(2n+3)} \geq s_t^{(2n+1)} \geq s_t^{(1)} \tag{11.5}$$

for all t, $0 \leq t \leq T$, and for all $n \geq n_0$. Moreover, there exists a positive integer n_0 such that

$$s_t^{(2n)} = s_t^{(2n_0)} = s_t^{even} \geq s_t^{odd} = s_t^{(2n_0+1)} = s_t^{(2n+1)} \tag{11.6}$$

for all t, $0 \leq t \leq T$, and for all $n \geq n_0$.

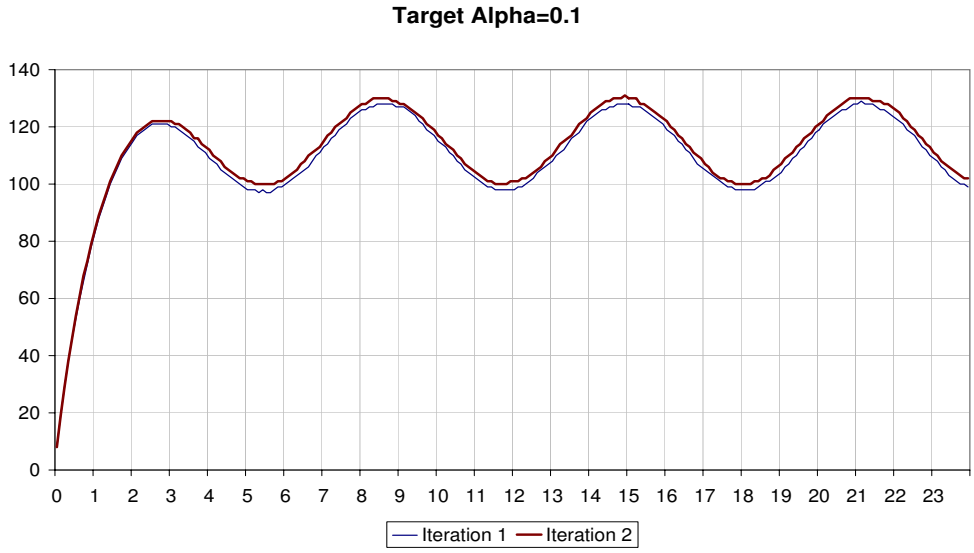
Proof. Given that $s_t^{(0)} = \infty$, we necessarily have $s^(0)_t > s_t^{(1)}$ for all t, $0 \leq t \leq T$. Hence we have the ordering of the initial ordering of the staffing functions that lets us apply the stochastic order. We then proceed recursively. As a consequence of the sample-path stochastic order, we get ordinary stochastic order in (11.3), we get ordinary stochastic order $N_t^{(1)} \leq_{st} N_t^{(2)}$ for all t. Ordinary stochastic order is equivalent to the tail probabilities being ordered: $P(N_t^{(1)} > x) \leq P(N_t^{(2)} > x)$ for all x, which implies the ordering for the staffing functions at time t. In particular, suppose that

$$P\left(N_t^{(2)} \geq s_t^{(2)}\right) \leq \alpha < P\left(N_t^{(2)} \geq s_t^{(2)} - 1\right) \ .$$

Since $P\left(N_t^{(1)} \geq s_t^{(2)}\right) \leq P\left(N_t^{(2)} \geq s_t^{(2)}\right) \leq \alpha$, necessarily $s_t^{(1)} \leq s_t^{(2)}$.

Case 1: $r > 1$**.** For $s_t^{(0)} = \infty$, we necessarily start with $s_t^{(0)} > s_t^{(1)}$ for all t, which produces first $N_t^{(1)} \leq_{st} N_t^{(0)}$ and then $s_t^{(2)} \leq s_t^{(1)}$ for all t. Continuing, we get $N_t^{(n)}$ stochastically decreasing in n and $s_t^{(n)}$ decreasing in n, again for all t. Since the staffing levels are integers, if we use only finitely many values of t, as in our implementation, then we necessarily get convergence in finitely many steps.

Figure 46: **Algorithm dynamics: range of staffing level for target α=0.1**

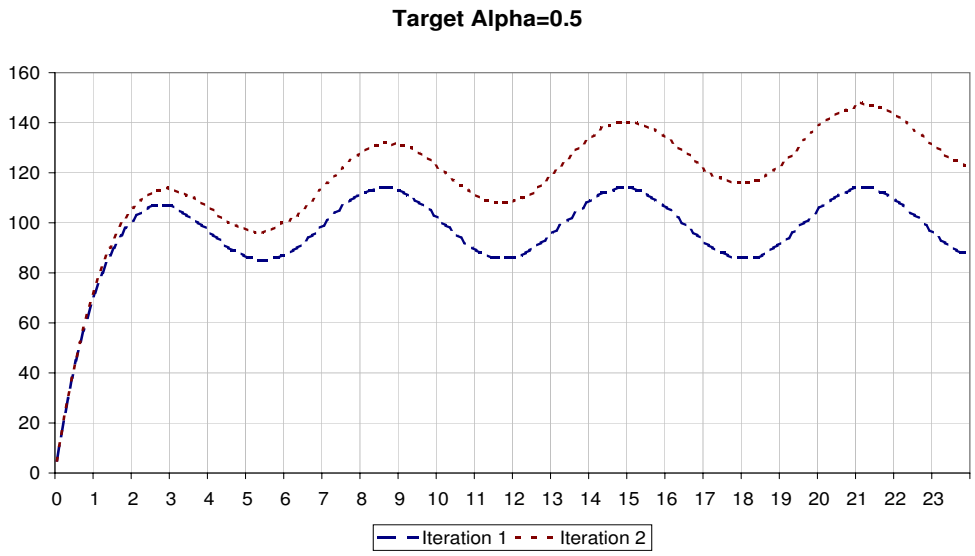


Case 2: $r < 1$. For $s_t^{(0)} = \infty$, we again necessarily start with $s_t^{(0)} > s_t^{(1)}$ for all t. That produces first $N_t^{(1)} \geq_{st} N_t^{(0)}$ and then $s_t^{(0)} \geq s_t^{(2)} \geq s_t^{(1)}$ for all t. Afterwards, we get $N_t^{(1)} \geq_{st} N_t^{(2)} \geq_{st} N_t^{(0)}$ and $s_t^{(0)} \geq s_t^{(2)} \geq s_t^{(3)} \geq s_t^{(1)}$ for all t. Continuing, we get $N_t^{(2n)}$ stochastically increasing in n, while $N_t^{(2n+1)}$ stochastically decreases in n, for all t. Similarly, $s_t^{(2n)}$ decreases in n, while $s_t^{(2n+1)}$ increases in n for all t. We thus have convergence, to possibly different limits. Since the staffing levels are integers, if we use only finitely many values of t, as in our implementation, then we necessarily get convergence in finitely many steps. ∎

We remark that we also obtain the convergence in Theorem 11.2 with other initial conditions. In particular, it suffices to let $s_t^{(0)}$ be sufficiently large for all t. For $r > 1$, it suffices to have $s_t^{(0)} \geq s_t^{ISA}$ for all t. For $r < 1$, it suffices to have $s_t^{(0)} \geq s_t^{even}$ for all t.

We conclude this section by making some empirical observations, for which we have yet to develop supporting theory. We also observed that the target delay probability α strongly influenced the dynamics. In particular, higher values of α cause larger oscillations in the oscillating case, and slower convergence to the limit in all cases.

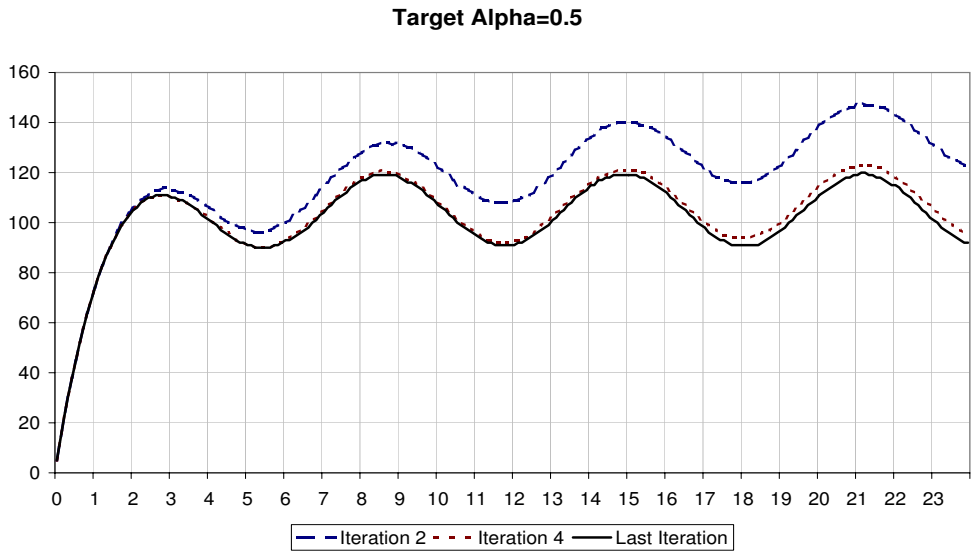
Figure 47: **Algorithm dynamics: range of staffing level for target α=0.5**



11.4. Convergence First at Smaller Times

Finally, we also observed a time-dependent behavior in the convergence of $s_t^{(n)}$. We observed a greater gap as time increased. For example, let $I_t \equiv \inf\{j : s_t^{(i)} = s_t^{(j)} \quad \text{for all} \quad i \geq j\}$. We observed that $I_{t_2} \geq I_{t_1}$ for all $t_2 > t_1$. An illustration can be viewed in Figures 45, 47 and 48. This time-dependent behavior is understandable, because the gap between two different staffing levels persists across time, so that there is a gap in the death rates at each t. Hence, as t gets larger, the two processes can get further apart. Thus the gap can first decrease more at the initial times. When it reaches the limit at earlier times, the gap will still have to decrease more at later times.

Figure 48: **Algorithm dynamics: evolution of convergence during algorithm runtime**



12. An Asymptotic Perspective

We can create a rigorous framework for the square-root-staffing formula by applying the asymptotic analysis of uniform acceleration to multi-server queues with abandonment. The underlying intuition for optimal staffing is that for large systems we staff exactly for the number of customers requesting service so as a first order effect, abandonment simply does not happen. Thus the associated fluid model should not be a function of any abandonment parameters. The effects of abandonment appear as second order phenomena at best and are found in the associated diffusion model. Moreover, we can show that for the special case of $\theta = \mu$, our limiting diffusion gives us exactly the square-root-staffing formula.

12.1. Limits for a Family of Multi-Server Queues with Abandonment

In this section we will consider a family of Markovian $M_t/M/s_t + M$ models indexed by a parameter η. As before, we will focus on the stochastic process representing the number of customers in the system, which is a birth-and-death process. We will identify that stochastic process with the $M_t/M/s_t + M$ model.

For each $\eta > 0$, let Let $\{ N^\eta \mid \eta > 0 \}$ by a family of multi-server queues with abandonment indexed by η, where $\theta^\eta = \theta$ and $\mu^\eta = \mu$ (i.e., the service and abandonment rates are independent of η), but

$$\lambda_t^\eta = \eta \cdot \lambda_t \quad \text{and} \quad s_t^\eta = \eta \cdot s_t^{(f)} + \sqrt{\eta} \cdot s_t^{(d)} + o(\sqrt{\eta}\,). \tag{12.1}$$

(The superscripts f and d on $s_t^{(d)}$ and $s_t^{(d)}$ indicate the "fluid-approximation" term and the "diffusion-approximation" term, respectively.)

Unlike the uniform acceleration scalings that lead to the pointwise stationary approximation, as in Massey and Whitt (1998), this one is inspired by the scalings of Halfin and Whitt (1981), Garnett et al. (2002) and Mandelbaum, Massey and Reiman (1998). Here we are scaling up the arrival rate (representing "demand" for our call center service) and the number of service agents (representing "supply" for our call center service) by the same parameter η. By limit theorems developed in Mandelbaum, Massey and Reiman (1998), we know that such a family of processes have fluid and diffusion approximations as $\eta \to \infty$. We want to restrict ourselves to a special type of growth behavior for the number of servers.

Theorem 12.1. *Consider the family of multiserver queues with abandonment having the growth conditions for its parameters as defined above. If we set*

$$s_t^\eta = \eta \cdot m_t + \sqrt{\eta} \cdot s_t^{(d)} + o(\sqrt{\eta}\,) \tag{12.2}$$

i.e., if we use (12.1) *with* $s_t^{(f)} = m_t$, *where*

$$\frac{d}{dt}m_t = \lambda_t - \mu_t \cdot m_t, \tag{12.3}$$

then

$$\lim_{\eta\to\infty} P\left(N_t^\eta \geq s_t^\eta\right) = p\left(N_t^{(d)} \geq s_t^{(d)}\right), \tag{12.4}$$

where $N^{(d)} = \left\{ N_t^{(d)} \mid t \geq 0 \right\}$ *is a diffusion process, which is the unique sample-path solution to the integral equation*

$$\begin{aligned} N_t^{(d)} &= N_0^{(d)} + \int_0^t (\mu_u - \theta_u) \cdot (s_u^{(d)})^- du \\ &\quad - \int_0^t \left(\theta_u \cdot (N_u^{(d)})^+ - \mu_u \cdot (N_u^{(d)})^-\right) du + B\left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du\right) \end{aligned} \tag{12.5}$$

and the process $\{ B(t) \mid t \geq 0 \}$ *is standard Brownian motion.*

Thus we can reduce the analysis of the probability of delay (approximately) to the analysis of a one-dimensional diffusion $N^{(d)}$. Notice that since λ_t and μ_t are given, then so is m_t. Thus server staffing for this model can only be controlled by the selection of $s^{(d)}$. Also notice that the diffusion $N^{(d)}$ is independent of $s^{(d)}$ *as long as* $\theta_t = \mu_t$ or $s_t^{(d)} \geq 0$ for all time $t \geq 0$.

For the special case of $\mu = \theta$ we can give a complete analysis of the delay probabilities that gives the MOL server-staffing heuristic.

Corollary 12.1. *If* $\theta = \mu$ *and* $s_t^\eta = \eta \cdot m_t + \Phi^{-1}(1-\alpha) \cdot \sqrt{\eta \cdot m_t}$, *where*

$$\frac{1}{\sqrt{2\pi}} \int_{\Phi^{-1}(1-\alpha)}^{\infty} e^{-x^2/2} dx = \alpha, \tag{12.6}$$

then we have

$$\lim_{\eta\to\infty} (N_t^\eta \geq s_t^\eta) = \alpha \tag{12.7}$$

for all $t > 0$.

Unfortunately, $N^{(d)}$ in general is *not* a Gaussian process. This also means that the following set of differential equations are not autonomous.

Corollary 12.2. *The differential equation for the mean of* $N^{(d)}$ *is*

$$\frac{d}{dt}E\left[N_t^{(d)}\right] = (\mu_t - \theta_t) \cdot (s_t^{(d)})^- - \theta_t \cdot E\left[(N_t^{(d)})^+\right] + \mu_t \cdot E\left[(N_t^{(d)})^-\right]. \tag{12.8}$$

Since $(N_t^{(d)})^+ \cdot (N_t^{(d)})^- = 0$, *the differential equation for the variance of* $N^{(d)}$ *equals*

$$\begin{aligned} \frac{d}{dt}\mathsf{Var}\left[N_t^{(d)}\right] &= -2\theta_t \cdot \mathsf{Var}\left[(N_t^{(d)})^+\right] - 2\mu_t \cdot \mathsf{Var}\left[(N_t^{(d)})^-\right] \\ &\quad -2(\theta_t + \mu_t) \cdot E\left[(N_t^{(d)})^+\right] \cdot E\left[(N_t^{(d)})^-\right] + \lambda_t + \mu_t \cdot m_t. \end{aligned} \tag{12.9}$$

Proof of Theorem 12.1 . Define the function $f_t^\eta(\cdot)$, where

$$f_t^\eta(x) = \eta \cdot \lambda_t - \theta_t \cdot (\eta \cdot x - s_t^\eta)^+ - \mu_t \cdot (\eta \cdot x \wedge s_t^\eta). \tag{12.10}$$

Now we have

$$\begin{aligned} f_t^\eta(x) &= \eta \cdot \lambda_t - \theta_t \cdot (\eta x - s_t^\eta)^+ - \mu_t \cdot ((\eta x) \wedge s_t^\eta) \\ &= \eta \cdot \lambda_t - \eta \cdot \theta_t \cdot x + (\theta_t - \mu_t) \cdot ((\eta \cdot x) \wedge s_t^\eta)\,. \end{aligned}$$

However

$$\begin{aligned} (\eta \cdot x) \wedge s_t^\eta &= (\eta \cdot x) \wedge \left(\eta \cdot m_t + \sqrt{\eta} \cdot s_t^{(d)} + o(\sqrt{\eta}\,)\right) \\ &= 1_{\{x<m_t\}} \cdot (\eta \cdot x + o(\sqrt{\eta}\,)) + 1_{\{x=m_t\}} \cdot (\eta \cdot m_t - \sqrt{\eta} \cdot (s_t^{(d)})^- + o(\sqrt{\eta}\,)) \\ &\quad + 1_{\{x>m_t\}} \cdot (\eta \cdot m_t - \sqrt{\eta} \cdot s_t^{(d)} + o(\sqrt{\eta}\,)) \\ &= \eta \cdot (x \wedge m_t) + \sqrt{\eta} \cdot \left((s_t^{(d)})^+ 1_{\{x>m_t\}} - (s_t^{(d)})^- 1_{\{x \geq m_t\}}\right) + o(\sqrt{\eta}\,) \end{aligned}$$

combining these results gives us the asymptotic expansion

$$\begin{aligned} f_t^\eta(x) &= \eta \cdot \left(\lambda_t - \theta_t \cdot (x - m_t)^+ - \mu_t \cdot (x \wedge m_t)\right) \\ &\quad + \sqrt{\eta} \cdot (\theta_t - \mu_t) \left((s_t^{(d)})^+ \cdot 1_{\{x>m_t\}} - (s_t^{(d)})^- \cdot 1_{\{x \geq m_t\}}\right) + o(\sqrt{\eta}\,) \end{aligned}$$

as $\eta \to \infty$.

It follows that $f_t^\eta = \eta \cdot f_t^{(f)} + \sqrt{\eta} \cdot f_t^{(d)} + o(\sqrt{\eta}\,)$, where

$$f_t^{(f)}(x) = \lambda_t - \theta_t \cdot (x - m_t)^+ - \mu_t \cdot (x \wedge m_t) \tag{12.11}$$

and

$$f_t^{(d)}(x) = (\theta_t - \mu_t) \cdot \left((s_t^{(d)})^+ \cdot 1_{\{x>m_t\}} - (s_t^{(d)})^- \cdot 1_{\{x \geq m_t\}}\right). \tag{12.12}$$

Now

$$\Lambda f_t^{(f)}(x; y) = (\theta_t - \mu_t) \cdot \left(y \cdot 1_{\{x<m_t\}} - y^- \cdot 1_{\{x=m_t\}}\right) - \theta_t \cdot y\ , \tag{12.13}$$

where $\Lambda g(x; y) = g'(x+)y^+ - g'(x-)y^-$ is the *non-smooth derivative* of any function g that has left and right derivatives. Hence we have

$$\Lambda f_t^{(f)}(m_t; y) = \mu_t \cdot y^- - \theta_t \cdot y^+ \ \text{ and } \ f_t^{(d)}(m_t) = (\mu_t - \theta_t)(s_t^{(1)})^- \tag{12.14}$$

Finally, we have

$$N_t^{(d)} = N_0^{(d)} + \int_0^t \left(\Lambda f_t^{(f)}\left(m_u; N_u^{(d)}\right) + f_t^{(d)}\left(m_u\right)\right) du \tag{12.15}$$

$$
\begin{aligned}
&+B\left(\int_0^t(\lambda_u+\mu_u\cdot m_u)du\right)\\
=\;& N_0^{(d)}-\int_0^t\left(\theta_u\cdot((N_u^{(d)})^+ + (s_u^{(d)})^-)-\mu_u\cdot((N_u^{(d)})^- + (s_u^{(d)})^-)\right)du\\
&+B\left(\int_0^t(\lambda_u+\mu_u\cdot m_u)du\right). \qquad (12.16)
\end{aligned}
$$

12.2. Case 1: $\theta_t = \mu_t$ for all t

We then have

$$N_t^{(d)} = N_0^{(d)} - \int_0^t \mu_u \cdot N_u^{(d)} du + B\left(\int_0^t(\lambda_u+\mu_u\cdot m_u)du\right). \qquad (12.17)$$

It follows that $N^{(d)}$ is a zero-mean Gaussian process (if $N_0^{(d)}=0$) and

$$\frac{d}{dt}\mathsf{Var}\left[N_t^{(d)}\right] = -2\mu_t\cdot\mathsf{Var}\left[N_t^{(d)}\right]+\lambda_t+\mu_t\cdot m_t. \qquad (12.18)$$

Moreover, if $m_0 = \mathsf{Var}\left[N_0^{(d)}\right]$, then $\mathsf{Var}\left[N_t^{(d)}\right] = m_t$ for all $t \geq 0$.

We remark that the simplification in this special case is to be expected, because we know from Section 10 that the $M_t/M_t/s_t + M_t$ model in this case reduces to the infinite-server $M_t/M_t/\infty$ model, which in turn - by making a time change - can be transformed into a $M_t/M/\infty$ model, for which the time-dependent distribution is known to be Poisson for all t, with the mean m_t in (2.2).

12.3. Case 2: $\theta_t = 0$

We then have

$$N_t^{(d)} = N_0^{(d)} + \int_0^t \mu_u\cdot\left((N_u^{(d)})^- + (s_u^{(d)})^-\right)du + B\left(\int_0^t(\lambda_u+\mu_u\cdot m_u)du\right). \qquad (12.19)$$

with

$$\frac{d}{dt}E\left[N_t^{(d)}\right] = \mu_t\cdot\left(E\left[(N_t^{(d)})^-\right]+(s_t^{(d)})^-\right) \qquad (12.20)$$

and

$$\frac{d}{dt}\mathsf{Var}\left[N_t^{(d)}\right] = -2\mu_t\cdot\left(\mathsf{Var}\left[(N_t^{(d)})^-\right]+E\left[(N_t^{(d)})^+\right]\cdot E\left[(N_t^{(d)})^-\right]\right)+\lambda_t+\mu_t\cdot m_t. \qquad (12.21)$$

To conclude this section, we summarise the implications for our proposal to staffing at the offered load in the QED regime. Here is the implication: Asymptotically, controlling the delay for this queue with abandonment is a *second order* staffing effort (selecting $s_t^{(d)}$) whereas the leading order staffing level is satisfied by using the offered load. Moreover, for the special case

of the abandonment rate equaling the service rate, we can apply this argument to rigorously obtain the square-root staffing formula for the multi-server queue without abandonment. This is also the one case where the diffusion $N^{(d)}$ is Gaussian.

13. Summary and Directions for Future Research

13.1. Summary

We have developed a simulation-based algorithm - ISA - that generates staffing functions for which performance has been shown to be stable in the face of time-varying arrival rates for the $M_t/M/s_t + M$ model. The results have been found to be remarkably robust, applying to all forms of time variation in the arrival-rate function, with or without abandonment, covering the ED, QD and QED operational regimes. All experiments were done with nine target delay probabilities, ranging from $\alpha = 0.1$ (QD) to $\alpha = 0.9$ (ED). In §11 we proved that the ISA converges for the $M_t/M/s_t + M$ model.

In our simulation experiments, we found that ISA performs essentially the same as the modified-offered-load (MOL) approximation (reviewed in §2) with and without customer abandonment. Thus we provided additional support for MOL and the square-root-staffing formula in (2.7) based on it (using arrival rate λ_t^{MOL} in (2.8)). As we saw in §6, in many applications the MOL approximation is well approximated itself by lagged PSA and, in easy cases, by PSA itself. To implement the MOL approximation with abandonments, we applied many-server heavy-traffic limits from Garnett et al. (2002), which yield the Garnett function in (5.4); just as Jennings et al. (1996) applied applied many-server heavy-traffic limits from Halfin and Whitt (1981) without customer abandonment.

Finally, we found that the simple approach of *staffing to the offered load* is remarkably effective in the QED regime (when $\alpha = 0.5$). That was substantiated time and again by having the ISA staffing function s_t^{ISA} fall on top of the offered load m_t^{∞}, as in case 3 in Figure 3. Of course, abandonment plays an important role; the staffing is always above the offered load without abandonment. When the service times are short, the offered load m_t^{∞} may agree closely with the PSA offered load $m_t^{PSA} \equiv \lambda(t)E[S]$; then staffing to the offered load reduces to the *naive deterministic approximation*: staffing to the PSA offered load m_t^{PSA}. However, it is good to be careful, because even for the realistic example in §6, PSA performed significantly worse than ISA, MOL and lagged PSA.

13.2. Next Steps

There is much yet to be done. Here are some natural next-steps:

1. As discussed in Section 5, for the $M_t/M/s_t + M$ model, it remains to explore alternative staffing methods to achieve better time-stability of abandonment probabilities and expected waiting times, especially under heavy loads, but experience indicates that the delay probability is a good performance target.

2. A great advantage of ISA is its generality. However, it remains to explore the ISA for additional queueing systems. We already have had partial (successful) results for deterministic and log-normal service-time distributions. It remains to consider other service-time distributions for the same models; it remains to consider other models. Some other models to analyze appear in Mandelbaum et al. (1998), e.g., queues with retrials and priority classes. Of special interest for actual call centers are multi-class models with skill-based routing. For call centers, our ultimate goal is to treat realistic multi-server systems with multiple call types and skill-based routing (SBR), but that remains to be done. In that setting, it is natural to apply SBR methods for stationary models after using the MOL approximation in (2.8) for each call type at time t. Once we have reduced the problem to a stationary SBR model, we may be able to apply the staffing method in Wallace and Whitt (2005). Approaches based on these ideas remain to be investigated. With networks of queues, the MOL approach can be applied together with results for networks of infinite-server queues; see Massey and Whitt (1993).

3. We proved that ISA converges for the $M_t/M/s_t + M$ model and we observed that it usually does so quite quickly, but it remains to analyze convergence of the algorithm more generally. Even for the $M_t/M/s_t + M$ model, some of the phenomena have not yet been adequately explained.

4. For one special case - the one with $\theta = \mu$ - we have provided strong theoretical support for our methods in §10 and §12. In §12 we exploited the mathematical framework of service networks in Mandelbaum et. al.(1998). It would be nice to prove much more generally that, under proper scaling, the actual time-dependent probability of delay under ISA indeed converges to the specified target as scale increases.

14. Acknowledgments

The reported research was supported by Grant No. 2002112 from the United States–Israel Binational Science Foundation (BSF). Avishai Mandelbaum and William A. Massey were also

supported by NSF grant DMI-0323668, while Ward Whitt was also supported by NSF grant DMI-0457095.

References

[1] Bolotin, V. 1994. Telephone circuit holding-time distributions. In *Proceedings of the International Teletraffic Congress, ITC* 14, J. Labetoulle and J. W. Roberts (eds.), North-Holland, Amsterdam, 125-134.

[2] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Statist. Assoc.* 100, 36–50.

[3] Eick, S., Massey, W. A., Whitt., W. 1993a. The Physics of The $M_t/G/\infty$ Queue. *Operations Research*, **41**(4), 731-742.

[4] Eick, S., Massey, W. A., Whitt, W. 1993b. $M_t/G/\infty$ Queues with Sinusoidal Arrival Rates. *Management Science*, **39**(2), 241-252.

[5] Gans, N., Koole, G., Mandelbaum, A. 2003. Telephone Call Centers: Tutorial, Review and Research Prospects. *Manufacturing and Service Operations Management*, **5**(2), 79–141.

[6] Garnett, O., Mandelbaum, A., Reiman, M. I. 2002. Designing a Call Center with Impatient Customers. *Manufacturing and Service Operations Management*, **4**(3), 208–227.

[7] Green, L. V., Kolesar, P. J. 1991. The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals. *Management Science*, **37**(1), 84–97.

[8] Green, L. V., Kolesar, P. J., Soares, J. 2001. Improving the SIPP Approach For Staffing Service Systems That Have Cyclic Demand. *Operations Research*, **49**, 549–564.

[9] Green, L. V., Kolesar, P. J., Whitt, W. 2005. Coping with Time-Varying Demand when Setting Staffing Requirements for a Service System. *Production and Operations Management*, forthcoming. Available at: http://www.columbia.edu/∼ww2040/Coping.pdf

[10] Halfin, S., Whitt, W. 1981. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, **29**, 567–587.

[11] Ingolfsson, A. 2005. Modeling the $M(t)/M/s(t)$ Queue with an Exhaustive Discipline. Available at: $http://www.bus.ualberta.ca/aingolfsson/working_papers.htm$

[12] Jagerman, D. L. 1975. Nonstationary Blocking in Telephone Traffic. *Bell System Technical Journal*, **54**, 625–661.

[13] Jelenkovic P., Mandelbaum A., Momcilovic P. 2004. Heavy Traffic Limits for Queues with Many Deterministic Servers. *Queueing Systems*, **47**, 53–69.

[14] Jennings, O. B., Mandelbaum, A., Massey, W. A., Whitt, W. 1996. Server Staffing to Meet Time-Varying Demand. *Management Science*, **42**(10), 1383–1394.

[15] Kamae, T., Krengel, U., O'Brien, G. L. 1978. Stochastic Inequalities on Partially Ordered Spaces. *Annals of Probability* **5**, 899–912.

[16] Mandelbaum, A., Massey, W.A., Reiman, M. I. 1998. Strong Approximations for Markovian Service Networks. *Queueing Systems: Theory and Applications (QUESTA)*, **30**, 149–201.

[17] Massey, W. A., Parker, G. A., Whitt, W. 1996. Estimating the Parameters of a Nonhomogeneous Poisson Process with Linear Rate. *Telecommunication Systems*, **5**, 361–388.

[18] Massey, W. A., Whitt, W. 1993. Networks of Infinite-Server Queues with Nonstationary Poisson Input. *Queueing Systems* **13** (1), 183–250.

[19] Massey, W. A., Whitt, W. 1994. An Analysis of the Modified Offered Load Approximation for the Erlang Loss Model. *Annals of Applied Probability*, **4**, 1145–1160.

[20] Massey, W. A., Whitt, W. 1997. Peak Congestion in Multi-Server Service Systems with Slowly Varying Arrival Rates. *Queueing Systems*, **25**, 157–172.

[21] Massey, W. A., Whitt, W. 1998. Uniform Acceleration Expansions for Markov Chains with Time-Varying Rates. *Annals of Applied Probability*, **9** (4), 1130–1155.

[22] Müller, A., Stoyan, D. 2002. *Comparison Methods for Stochastic Models and Risks*, Wiley.

[23] Ross, S. M. 1990. *A Course in Simulation*, Macmillan.

[24] Ross, S. M. 1996. *Stochastic Processes*, second edition, Wiley.

[25] Ross, S. M. 2003. *Introduction to Probability Models*, eighth edition, Academic Press.

[26] Wallace, R. B., Whitt, W. 2005. A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing and Service Operations Management*, forthcoming.
Available at: http://www.columbia.edu/∼ww2040/recent.html

[27] Whitt, W. Comparing Counting Processes and Queues. 1981. *Advances in Applied Probability* **13** 207–220.

[28] Whitt, W. 1991. The Pointwise Stationary Approximation for $M_t/M_t/s$ Queues Is Asymptotically Correct as the rate Increases. *Management Science*, **37**(2), 307–314.

[29] Whitt, W. 1992. Understanding the Efficiency of Multi-Server Service Systems. *Management Science*, **38**, 708–723.

[30] Whitt, W. 2000. The Impact of a Heavy-Tailed Service-Time Distribution upon the M/GI/s Waiting-Time Distribution. *Queueing Systems*, **36**, 71–87.

[31] Whitt, W. 2004. Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments. *Management Science*, **50** (10) 1449–1461.

[32] Whitt, W. 2005. Engineering Solution of a Basic Call-Center Model. *Management Science*, **51**, 221–235.